#21

NOVEL NUCLEIC ACIDS AND POLYPEPTIDES

1. RELATED APPLICATIONS

This application is a continuation in part of copending U.S. Patent Application Nos. 09/898,888 filed July 3, 2001 (Hyseq docket no. 748CON2), which is a continuation of 09/340,623 filed June 28, 1999 (now abandoned, Hyseq docket no. 748CON1), which is a continuation of 09/205,070 filed Dec. 3, 1998 (now abandoned, Hyseq docket no. 748); copending U.S. Application No. 09/919,002 filed July 30, 2001 (Hyseq docket no. 752CON2), which is a continuation of 09/359,922 filed July 22, 1999 (Hyseq docket no. 752CON1), which is a continuation of 09/205,155 filed Dec. 3, 1998 (now abandoned, Hyseq docket no. 752); copending U.S. Application No. 09/905,059 filed July 12, 2001 (Hyseq docket no. 778CON1), which is a continuation of 09/347,127 filed Jul. 2, 1999 (now abandoned, Hyseq docket no. 778); copending U.S. Application No. 09/952,981 filed Sept. 14, 2001 (Hyseq docket no. 779CON1), which is a continuation of copending 09/457,877 filed Dec. 8, 1999 (Hyseq docket no. 779); copending U.S. Application No. 09/471,275 filed Dec. 23, 1999 (Hyseq docket no. 782); copending U.S. Application No. 09/552,317 filed Apr. 25, 2000 (Hyseq docket no.:784CIP); copending U.S. Application No. 09/488,725 filed Jan. 21, 2000 (Hyseq docket no. 784); copending U.S. Application No. 09/922,279 filed Aug. 3, 2001 (Hyseq docket no. 785CON1), which is a continuation of 09/491,404 filed Jan. 25, 2000 (now abandoned, Hyseq docket no. 785); copending U.S. Application No. 09/560,875 filed Apr. 27, 2000 (Hyseq docket no. 787CIP); copending U.S. Application No. 09/496,914 filed Feb. 3, 2000 (Hyseq docket no. 787); copending U.S. Application No. 09/577,409 filed May 18, 2000 (Hyseq docket no. 788CIP); copending U.S. Application No. 09/515,126 filed Feb. 28, 2000 (Hyseq docket no. 788); copending U.S. Application No. 09/574,454 filed May 19, 2000 (Hyseq docket no. 789CIP); copending U.S. Application No. 09/519,705 filed Mar. 7, 2000 (Hyseq docket no. 789); copending U.S. Application No. 09/649,167 filed Aug. 23, 2000 (Hyseq docket no. 790CIP); copending U.S. Application No. 09/540,217 filed Mar. 31, 2000 (Hyseq docket no. 790); copending U.S. Application No. 09/770,160 filed Jan. 26, 2001 (Hyseq docket no. 791CIP); copending U.S. Application No. 10/014,487 filed Nov. 8, 2001 (Hyseq docket no.:791CON), which is a continuation of copending U.S. Application No. 09/552,929 filed Apr. 18, 2000 (Hyseq docket no. 791); and copending U.S. Application No. 09/989,660 filed Nov. 21, 2001 (Hyseq docket no. 792CON), which is a continuation of copending 09/577,408 filed May 18, 2000 (Hyseq docket no. 792). The entirety of the aforementioned applications are hereby incorporated by reference.

2. TECHNICAL FIELD

The present invention provides novel polynucleotides and proteins encoded by such polynucleotides, along with uses for these polynucleotides and proteins, for example in therapeutic, diagnostic and research methods.

3. BACKGROUND

Technology aimed at the discovery of protein factors (including *e.g.*, cytokines, such as lymphokines, interferons, CSFs, chemokines, and interleukins) has matured rapidly over the past decade. The now routine hybridization cloning and expression cloning techniques clone novel polynucleotides "directly" in the sense that they rely on information directly related to the discovered protein (*i.e.*, partial DNA/amino acid sequence of the protein in the case of hybridization cloning; activity of the protein in the case of expression cloning). More recent "indirect" cloning techniques such as signal sequence cloning, which isolates DNA sequences based on the presence of a now well-recognized secretory leader sequence motif, as well as various PCR-based or low stringency hybridization-based cloning techniques, have advanced the state of the art by making available large numbers of DNA/amino acid sequences for proteins that are known to have biological activity, for example, by virtue of their secreted nature in the case of leader sequence cloning, by virtue of their cell or tissue source in the case of PCR-based techniques, or by virtue of structural similarity to other genes of known biological activity.

Identified polynucleotide and polypeptide sequences have numerous applications in, for example, diagnostics, forensics, gene mapping; identification of mutations responsible for genetic disorders or other traits, to assess biodiversity, and to produce many other types of data and products dependent on DNA and amino acid sequences.

4. SUMMARY OF THE INVENTION

The compositions of the present invention include novel isolated polypeptides, novel isolated polynucleotides encoding such polypeptides, including recombinant DNA molecules, cloned genes or degenerate variants thereof, especially naturally occurring

variants such as allelic variants, antisense polynucleotide molecules, and antibodies that specifically recognize one or more epitopes present on such polypeptides, as well as hybridomas producing such antibodies.

The compositions of the present invention additionally include vectors, including expression vectors, containing the polynucleotides of the invention, cells genetically engineered to contain such polynucleotides and cells genetically engineered to express such polynucleotides.

The present invention relates to a collection or library of novel nucleic acid sequences assembled from expressed sequence tags (ESTs) isolated mainly by sequencing by hybridization (SBH), and in some cases, sequences obtained from one or more public databases. The invention also relates to the proteins encoded by such polynucleotides, along with therapeutic, diagnostic and research utilities for these polynucleotides and proteins. These sequences of the present invention are designated herein as 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502. The sequences designated 748 SEQ ID NO: 1-45,207 correspond to SEQ ID NO: 1-45,196 in the Sequence Listing; sequences 752 SEQ ID NO: 1-13,203 correspond to SEQ ID NO: 45,208-58,410 in the Sequence Listing; sequences 778 SEQ ID NO: 1-105 correspond to SEQ ID NO: 58,411-58,515; sequences 779 SEQ ID NO: 1-128 correspond to SEQ ID NO: 58,516-58,643 in the Sequence Listing; sequences 782 SEQ ID NO: 1-10,451 correspond to SEQ ID NO: 58,664-69,094 in the Sequence Listing; sequences 784 SEQ ID NO: 1-10,289 correspond to SEQ ID NO: 69,095-79,383; sequences 785 SEQ ID NO: 1-3796 correspond to SEQ ID NO: 79,384-83,179 in the Sequence Listing; sequences 787 SEQ ID NO: 10,410 correspond to SEQ ID NO: 83,180-93,589 in the Sequence Listing; sequences 788 SEQ ID NO: 1-14,074 correspond to SEQ ID NO: 93,590-107,663 in the Sequence Listing; sequences 789 SEQ ID NO: 1-6391 correspond to SEQ ID NO: 107,664-114,054 in the Sequence Listing; sequences 790 SEQ ID NO: 1-30,533 correspond to SEQ ID NO: 114,055-144,607 in the Sequence Listing; sequences 791 SEQ ID NO: 1-5822 correspond to SEQ ID NO: 144,608-150,429 in the Sequence Listing; and sequences 792 SEQ ID NO: 1-8502 correspond to

SEQ ID NO: 150,430-158,931 in the Sequence Listing. The nucleic acids and polypeptides are provided in the Sequence Listing, wherein for the nucleic acids, A is adenosine; C is cytosine; G is guanine; T is thymine; and N is any of the four bases. In the amino acids provided in the Sequence Listing, * corresponds to the stop codon.

The nucleic acid sequences of the present invention also include, nucleic acid sequences that hybridize to the complement of the aforementioned nucleic acid sequences under stringent hybridization conditions; nucleic acid sequences which are allelic variants or species homologues of any of the nucleic acid sequences recited above, or nucleic acid sequences that encode a peptide comprising a specific domain or truncation of the peptides encoded by said nucleic acid sequences, a polynucleotide comprising a nucleotide sequence having at least 90% identity to an identifying sequence or a degenerate variant or fragment thereof. The identifying sequence can be 100 base pairs in length.

The nucleic acid sequences of the present invention also include the sequence information from the aforementioned nucleic acid sequences of 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502 . The sequence information can be a segment of any one of 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502 that uniquely identifies or represents the sequence information of 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502.

A collection as used in this application can be a collection of only one polynucleotide. The collection of sequence information or identifying information of each sequence can be provided on a nucleic acid array. In one embodiment, segments of sequence information is provided on a nucleic acid array to detect the polynucleotide that

contains the segment. The array can be designed to detect full-match or mismatch to the polynucleotide that contains the segment. The collection can also be provided in a computer-readable format.

This invention also includes the reverse or direct complement of any of the nucleic acid sequences recited above; cloning or expression vectors containing the nucleic acid sequences; and host cells or organisms transformed with these expression vectors. Nucleic acid sequences (or their reverse or direct complements) according to the invention have numerous applications in a variety of techniques known to those skilled in the art of molecular biology, such as use as hybridization probes, use as primers for PCR, use in an array, use in computer-readable media, use in sequencing full-length genes, use for chromosome and gene mapping, use in the recombinant production of protein, and use in the generation of anti-sense DNA or RNA, their chemical analogs and the like.

In a preferred embodiment, the nucleic acid sequences of 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502 or novel segments or parts of the nucleic acids of the invention are used as primers in expression assays that are well known in the art. In a particularly preferred embodiment, the nucleic acid sequences of 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502 or novel segments or parts of the nucleic acids provided herein are used in diagnostics for identifying expressed genes or, as well known in the art and exemplified by Vollrath *et al.*, Science 258:52-59 (1992), as expressed sequence tags for physical mapping of the human genome.

The isolated polynucleotides of the invention include, but are not limited to, a polynucleotide comprising any one of the nucleotide sequences set forth in 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-

30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502; a polynucleotide comprising any of the full length protein coding sequences of 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502; and a polynucleotide comprising any of the nucleotide sequences of the mature protein coding sequences of 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502. The polynucleotides of the present invention also include, but are not limited to, a polynucleotide that hybridizes under stringent hybridization conditions to (a) the complement of any one of the nucleotide sequences set forth in 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502; (b) a nucleotide sequence encoding any one of the amino acid sequences set forth in the Sequence Listing; (c) a polynucleotide which is an allelic variant of any polynucleotides recited above; (d) a polynucleotide which encodes a species homolog (e.g. orthologs) of any of the proteins recited above; or (e) a polynucleotide that encodes a polypeptide comprising a specific domain or truncation of any of the polypeptides comprising an amino acid sequence set forth in the Sequence Listing.

The isolated polypeptides of the invention include, but are not limited to, a polypeptide comprising any of the amino acid sequences set forth in the Sequence Listing; or the corresponding full length or mature protein. Polypeptides of the invention also include polypeptides with biological activity that are encoded by (a) any of the polynucleotides having a nucleotide sequence set forth in 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-

5822, and 792 SEQ ID NO: 1-8502; or (b) polynucleotides that hybridize to the complement of the polynucleotides of (a) under stringent hybridization conditions. Biologically or immunologically active variants of any of the polypeptide sequences in the Sequence Listing, and "substantial equivalents" thereof (*e.g.*, with at least about 65%, 70%, 75%, 80%, 85%, 90%, 95%, 98% or 99% amino acid sequence identity) that preferably retain biological activity are also contemplated. The polypeptides of the invention may be wholly or partially chemically synthesized but are preferably produced by recombinant means using the genetically engineered cells (e.g. host cells) of the invention.

The invention also provides compositions comprising a polypeptide of the invention. Polypeptide compositions of the invention may further comprise an acceptable carrier, such as a hydrophilic, *e.g.*, pharmaceutically acceptable, carrier.

The invention also provides host cells transformed or transfected with a polynucleotide of the invention.

The invention also relates to methods for producing a polypeptide of the invention comprising growing a culture of the host cells of the invention in a suitable culture medium under conditions permitting expression of the desired polypeptide, and purifying the polypeptide from the culture or from the host cells. Preferred embodiments include those in which the protein produced by such process is a mature form of the protein.

Polynucleotides according to the invention have numerous applications in a variety of techniques known to those skilled in the art of molecular biology. These techniques include use as hybridization probes, use as oligomers, or primers, for PCR, use for chromosome and gene mapping, use in the recombinant production of protein, and use in generation of anti-sense DNA or RNA, their chemical analogs and the like. For example, when the expression of an mRNA is largely restricted to a particular cell or tissue type, polynucleotides of the invention can be used as hybridization probes to detect the presence of the particular cell or tissue mRNA in a sample using, *e.g.*, *in situ* hybridization.

In other exemplary embodiments, the polynucleotides are used in diagnostics as expressed sequence tags for identifying expressed genes or, as well known in the art and exemplified by Vollrath *et al.*, Science 258:52-59 (1992), as expressed sequence tags for physical mapping of the human genome.

The polypeptides according to the invention can be used in a variety of conventional procedures and methods that are currently applied to other proteins. For example, a polypeptide of the invention can be used to generate an antibody that specifically binds the polypeptide. Such antibodies, particularly monoclonal antibodies, are useful for detecting or quantitating the polypeptide in tissue. The polypeptides of the invention can also be used as molecular weight markers, and as a food supplement.

Methods are also provided for preventing, treating, or ameliorating a medical condition which comprises the step of administering to a mammalian subject a therapeutically effective amount of a composition comprising a polypeptide of the present invention and a pharmaceutically acceptable carrier.

In particular, the polypeptides and polynucleotides of the invention can be utilized, for example, in methods for the prevention and/or treatment of disorders involving aberrant protein expression or biological activity.

The present invention further relates to methods for detecting the presence of the polynucleotides or polypeptides of the invention in a sample. Such methods can, for example, be utilized as part of prognostic and diagnostic evaluation of disorders as recited herein and for the identification of subjects exhibiting a predisposition to such conditions. The invention provides a method for detecting the polynucleotides of the invention in a sample, comprising contacting the sample with a compound that binds to and forms a complex with the polynucleotide of interest for a period sufficient to form the complex and under conditions sufficient to form a complex and detecting the complex such that if a complex is detected, the polynucleotide of interest is detected. The invention also provides a method for detecting the polypeptides of the invention in a sample comprising contacting the sample with a compound that binds to and forms a complex with the polypeptide under conditions and for a period sufficient to form the complex and detecting the formation of the complex such that if a complex is formed, the polypeptide is detected.

The invention also provides kits comprising polynucleotide probes and/or monoclonal antibodies, and optionally quantitative standards, for carrying out methods of the invention. Furthermore, the invention provides methods for evaluating the efficacy of

drugs, and monitoring the progress of patients, involved in clinical trials for the treatment of disorders as recited above.

The invention also provides methods for the identification of compounds that modulate (*i.e.*, increase or decrease) the expression or activity of the polynucleotides and/or polypeptides of the invention. Such methods can be utilized, for example, for the identification of compounds that can ameliorate symptoms of disorders as recited herein. Such methods can include, but are not limited to, assays for identifying compounds and other substances that interact with (*e.g.*, bind to) the polypeptides of the invention. The invention provides a method for identifying a compound that binds to the polypeptides of the invention comprising contacting the compound with a polypeptide of the invention in a cell for a time sufficient to form a polypeptide/compound complex, wherein the complex drives expression of a reporter gene sequence in the cell; and detecting the complex by detecting the reporter gene sequence expression such that if expression of the reporter gene is detected the compound the binds to a polypeptide of the invention is identified.

The methods of the invention also provides methods for treatment which involve the administration of the polynucleotides or polypeptides of the invention to individuals exhibiting symptoms or tendencies. In addition, the invention encompasses methods for treating diseases or disorders as recited herein comprising administering compounds and other substances that modulate the overall activity of the target gene products. Compounds and other substances can effect such modulation either on the level of target gene/protein expression or target protein activity.

The polypeptides of the present invention and the polynucleotides encoding them are also useful for the same functions known to one of skill in the art as the polypeptides and polynucleotides to which they have homology. If no homology is set forth for a sequence, then the polypeptides and polynucleotides of the present invention are useful for a variety of applications, as described herein, including use in arrays for detection.

5. DETAILED DESCRIPTION OF THE INVENTION

5.1 DEFINITIONS

It must be noted that as used herein and in the appended claims, the singular forms "a", "an" and "the" include plural references unless the context clearly dictates otherwise.

The term "active" refers to those forms of the polypeptide which retain the biologic and/or immunologic activities of any naturally occurring polypeptide. According to the invention, the terms "biologically active" or "biological activity" refer to a protein or peptide having structural, regulatory or biochemical functions of a naturally occurring molecule. Likewise "immunologically active" or "immunological activity" refers to the capability of the natural, recombinant or synthetic polypeptide to induce a specific immune response in appropriate animals or cells and to bind with specific antibodies.

The term "activated cells" as used in this application are those cells which are engaged in extracellular or intracellular membrane trafficking, including the export of secretory or enzymatic molecules as part of a normal or disease process.

The terms "complementary" or "complementarity" refer to the natural binding of polynucleotides by base pairing. For example, the sequence 5'-AGT-3' binds to the complementary sequence 3'-TCA-5'. Complementarity between two single-stranded molecules may be "partial" such that only some of the nucleic acids bind or it may be "complete" such that total complementarity exists between the single stranded molecules. The degree of complementarity between the nucleic acid strands has significant effects on the efficiency and strength of the hybridization between the nucleic acid strands.

The term "embryonic stem cells (ES)" refers to a cell that can give rise to many differentiated cell types in an embryo or an adult, including the germ cells. The term "germ line stem cells (GSCs)" refers to stem cells derived from primordial stem cells that provide a steady and continuous source of germ cells for the production of gametes. The term "primordial germ cells (PGCs)" refers to a small population of cells set aside from other cell lineages particularly from the yolk sac, mesenteries, or gonadal ridges during embryogenesis that have the potential to differentiate into germ cells and other cells. PGCs are the source from which GSCs and ES cells are derived The PGCs, the GSCs and the ES cells are capable of self-renewal. Thus these cells not only populate the germ

line and give rise to a plurality of terminally differentiated cells that comprise the adult specialized organs, but are able to regenerate themselves.

The term "expression modulating fragment," EMF, means a series of nucleotides which modulates the expression of an operably linked ORF or another EMF.

As used herein, a sequence is said to "modulate the expression of an operably linked sequence" when the expression of the sequence is altered by the presence of the EMF. EMFs include, but are not limited to, promoters, and promoter modulating sequences (inducible elements). One class of EMFs are nucleic acid fragments which induce the expression of an operably linked ORF in response to a specific regulatory factor or physiological event.

The terms "nucleotide sequence" or "nucleic acid" or "polynucleotide" or "oligonculeotide" are used interchangeably and refer to a heteropolymer of nucleotides or the sequence of these nucleotides. These phrases also refer to DNA or RNA of genomic or synthetic origin which may be single-stranded or double-stranded and may represent the sense or the antisense strand, to peptide nucleic acid (PNA) or to any DNA-like or RNA-like material. In the sequences herein A is adenine, C is cytosine, T is thymine, G is guanine and N is A, C, G or T (U). It is contemplated that where the polynucleotide is RNA, the T (thymine) in the sequences provided herein is substituted with U (uracil). Generally, nucleic acid segments provided by this invention may be assembled from fragments of the genome and short oligonucleotide linkers, or from a series of oligonucleotides, or from individual nucleotides, to provide a synthetic nucleic acid which is capable of being expressed in a recombinant transcriptional unit comprising regulatory elements derived from a microbial or viral operon, or a eukaryotic gene.

The terms "oligonucleotide fragment" or a "polynucleotide fragment", "portion," or "segment" or "probe" or "primer" are used interchangeably and refer to a sequence of nucleotide residues which are at least about 5 nucleotides, more preferably at least about 7 nucleotides, more preferably at least about 9 nucleotides, more preferably at least about 11 nucleotides and most preferably at least about 17 nucleotides. The fragment is preferably less than about 500 nucleotides, preferably less than about 200 nucleotides, more preferably less than about 100 nucleotides, more preferably less than about 50 nucleotides and most preferably less than 30 nucleotides. Preferably the probe is from

about 6 nucleotides to about 200 nucleotides, preferably from about 15 to about 50 nucleotides, more preferably from about 17 to 30 nucleotides and most preferably from about 20 to 25 nucleotides. Preferably the fragments can be used in polymerase chain reaction (PCR), various hybridization procedures or microarray procedures to identify or amplify identical or related parts of mRNA or DNA molecules. A fragment or segment may uniquely identify each polynucleotide sequence of the present invention. Preferably the fragment comprises a sequence substantially similar to any one of 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502.

Probes may, for example, be used to determine whether specific mRNA molecules are present in a cell or tissue or to isolate similar nucleic acid sequences from chromosomal DNA as described by Walsh et. al., (Walsh, P.S. *et al.*, 1992, PCR Methods Appl 1:241-250). They may be labeled by nick translation, Klenow fill-in reaction, PCR, or other methods well known in the art. Probes of the present invention, their preparation and/or labeling are elaborated in Sambrook, J. *et al.*, 1989, Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory, NY; or Ausubel, F.M. *et al.*, 1989, Current Protocols in Molecular Biology, John Wiley & Sons, New York NY, both of which are incorporated herein by reference in their entirety.

The nucleic acid sequences of the present invention also include the sequence information from the nucleic acid sequences of 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502. The sequence information can be a segment of any one of 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502 that uniquely identifies or represents the sequence

information of that sequence of 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502. One such segment can be a twenty-mer nucleic acid sequence because the probability that a twenty-mer is fully matched in the human genome is 1 in 300. In the human genome, there are three billion base pairs in one set of chromosomes. Because 4^{20} possible twenty-mers exist, there are 300 times more twenty-mers than there are base pairs in a set of human chromosomes. Using the same analysis, the probability for a seventeen-mer to be fully matched in the human genome is approximately 1 in 5. When these segments are used in arrays for expression studies, fifteen-mer segments can be used. The probability that the fifteen-mer is fully matched in the expressed sequences is also approximately one in five because expressed sequences comprise less than approximately 5% of the entire genome sequence.

Similarly, when using sequence information for detecting a single mismatch, a segment can be a twenty-five mer. The probability that the twenty-five mer would appear in a human genome with a single mismatch is calculated by multiplying the probability for a full match ($1 \div 4^{25}$) times the increased probability for mismatch at each nucleotide position (3 x 25). The probability that an eighteen mer with a single mismatch can be detected in an array for expression studies is approximately one in five. The probability that a twenty-mer with a single mismatch can be detected in a human genome is approximately one in five.

The term "open reading frame," ORF, means a series of nucleotide triplets coding for amino acids without any termination codons and is a sequence translatable into protein.

The terms "operably linked" or "operably associated" refer to functionally related nucleic acid sequences. For example, a promoter is operably associated or operably linked with a coding sequence if the promoter controls the transcription of the coding sequence. While operably linked nucleic acid sequences can be contiguous and in the same reading frame, certain genetic elements e.g. repressor genes are not contiguously linked to the coding sequence but still control transcription/translation of the coding sequence.

The term "pluripotent" refers to the capability of a cell to differentiate into a number of differentiated cell types that are present in an adult organism. A pluripotent cell is restricted in its differentiation capability in comparison to a totipotent cell.

The terms "polypeptide" or "peptide" or "amino acid sequence" refer to an oligopeptide, peptide, polypeptide or protein sequence or fragment thereof and to naturally occurring or synthetic molecules. A polypeptide "fragment," "portion," or "segment" is a stretch of amino acid residues of at least about 5 amino acids, preferably at least about 7 amino acids, more preferably at least about 9 amino acids and most preferably at least about 17 or more amino acids. The peptide preferably is not greater than about 200 amino acids, more preferably less than 150 amino acids and most preferably less than 100 amino acids. Preferably the peptide is from about 5 to about 200 amino acids. To be active, any polypeptide must have sufficient length to display biological and/or immunological activity.

The term "naturally occurring polypeptide" refers to polypeptides produced by cells that have not been genetically engineered and specifically contemplates various polypeptides arising from post-translational modifications of the polypeptide including, but not limited to, acetylation, carboxylation, glycosylation, phosphorylation, lipidation and acylation.

The term "translated protein coding portion" means a sequence which encodes for the full length protein which may include any leader sequence or any processing sequence.

The term "mature protein coding sequence" means a sequence which encodes a peptide or protein without a signal or leader sequence. The "mature protein portion" means that portion of the protein which does not include a signal or leader sequence. The peptide may have been produced by processing in the cell which removes any leader/signal sequence. The mature protein portion may or may not include the initial methionine residue. The methionine residue may be removed from the protein during processing in the cell. The peptide may be produced synthetically or the protein may have been produced using a polynucleotide only encoding for the mature protein coding sequence.

The term "derivative" refers to polypeptides chemically modified by such techniques as ubiquitination, labeling (*e.g.,* with radionuclides or various enzymes), covalent polymer attachment such as pegylation (derivatization with polyethylene glycol) and insertion or substitution by chemical synthesis of amino acids such as ornithine, which do not normally occur in human proteins.

The term "variant"(or "analog") refers to any polypeptide differing from naturally occurring polypeptides by amino acid insertions, deletions, and substitutions, created using, *e g.*, recombinant DNA techniques. Guidance in determining which amino acid residues may be replaced, added or deleted without abolishing activities of interest, may be found by comparing the sequence of the particular polypeptide with that of homologous peptides and minimizing the number of amino acid sequence changes made in regions of high homology (conserved regions) or by replacing amino acids with consensus sequence.

Alternatively, recombinant variants encoding these same or similar polypeptides may be synthesized or selected by making use of the "redundancy" in the genetic code. Various codon substitutions, such as the silent changes which produce various restriction sites, may be introduced to optimize cloning into a plasmid or viral vector or expression in a particular prokaryotic or eukaryotic system. Mutations in the polynucleotide sequence may be reflected in the polypeptide or domains of other peptides added to the polypeptide to modify the properties of any part of the polypeptide, to change characteristics such as ligand-binding affinities, interchain affinities, or degradation/turnover rate.

Preferably, amino acid "substitutions" are the result of replacing one amino acid with another amino acid having similar structural and/or chemical properties, *i.e.,* conservative amino acid replacements. "Conservative" amino acid substitutions may be made on the basis of similarity in polarity, charge, solubility, hydrophobicity, hydrophilicity, and/or the amphipathic nature of the residues involved. For example, nonpolar (hydrophobic) amino acids include alanine, leucine, isoleucine, valine, proline, phenylalanine, tryptophan, and methionine; polar neutral amino acids include glycine, serine, threonine, cysteine, tyrosine, asparagine, and glutamine; positively charged (basic) amino acids include arginine, lysine, and histidine; and negatively charged (acidic) amino

acids include aspartic acid and glutamic acid. "Insertions" or "deletions" are preferably in the range of about 1 to 20 amino acids, more preferably 1 to 10 amino acids. The variation allowed may be experimentally determined by systematically making insertions, deletions, or substitutions of amino acids in a polypeptide molecule using recombinant DNA techniques and assaying the resulting recombinant variants for activity.

Alternatively, where alteration of function is desired, insertions, deletions or non-conservative alterations can be engineered to produce altered polypeptides. Such alterations can, for example, alter one or more of the biological functions or biochemical characteristics of the polypeptides of the invention. For example, such alterations may change polypeptide characteristics such as ligand-binding affinities, interchain affinities, or degradation/turnover rate. Further, such alterations can be selected so as to generate polypeptides that are better suited for expression, scale up and the like in the host cells chosen for expression. For example, cysteine residues can be deleted or substituted with another amino acid residue in order to eliminate disulfide bridges.

The terms "purified" or "substantially purified" as used herein denotes that the indicated nucleic acid or polypeptide is present in the substantial absence of other biological macromolecules, *e.g.*, polynucleotides, proteins, and the like. In one embodiment, the polynucleotide or polypeptide is purified such that it constitutes at least 95% by weight, more preferably at least 99% by weight, of the indicated biological macromolecules present (but water, buffers, and other small molecules, especially molecules having a molecular weight of less than 1000 daltons, can be present).

The term "isolated" as used herein refers to a nucleic acid or polypeptide separated from at least one other component (*e.g.*, nucleic acid or polypeptide) present with the nucleic acid or polypeptide in its natural source. In one embodiment, the nucleic acid or polypeptide is found in the presence of (if anything) only a solvent, buffer, ion, or other component normally present in a solution of the same. The terms "isolated" and "purified" do not encompass nucleic acids or polypeptides present in their natural source.

The term "recombinant," when used herein to refer to a polypeptide or protein, means that a polypeptide or protein is derived from recombinant (*e.g.*, microbial, insect, or mammalian) expression systems. "Microbial" refers to recombinant polypeptides or proteins made in bacterial or fungal (*e.g.*, yeast) expression systems. As a product,

"recombinant microbial" defines a polypeptide or protein essentially free of native endogenous substances and unaccompanied by associated native glycosylation. Polypeptides or proteins expressed in most bacterial cultures, *e.g.*, *E. coli*, will be free of glycosylation modifications; polypeptides or proteins expressed in yeast will have a glycosylation pattern in general different from those expressed in mammalian cells.

The term "recombinant expression vehicle or vector" refers to a plasmid or phage or virus or vector, for expressing a polypeptide from a DNA (RNA) sequence. An expression vehicle can comprise a transcriptional unit comprising an assembly of (1) a genetic element or elements having a regulatory role in gene expression, for example, promoters or enhancers, (2) a structural or coding sequence which is transcribed into mRNA and translated into protein, and (3) appropriate transcription initiation and termination sequences. Structural units intended for use in yeast or eukaryotic expression systems preferably include a leader sequence enabling extracellular secretion of translated protein by a host cell. Alternatively, where recombinant protein is expressed without a leader or transport sequence, it may include an amino terminal methionine residue. This residue may or may not be subsequently cleaved from the expressed recombinant protein to provide a final product.

The term "recombinant expression system" means host cells which have stably integrated a recombinant transcriptional unit into chromosomal DNA or carry the recombinant transcriptional unit extrachromosomally. Recombinant expression systems as defined herein will express heterologous polypeptides or proteins upon induction of the regulatory elements linked to the DNA segment or synthetic gene to be expressed. This term also means host cells which have stably integrated a recombinant genetic element or elements having a regulatory role in gene expression, for example, promoters or enhancers. Recombinant expression systems as defined herein will express polypeptides or proteins endogenous to the cell upon induction of the regulatory elements linked to the endogenous DNA segment or gene to be expressed. The cells can be prokaryotic or eukaryotic.

The term "secreted" includes a protein that is transported across or through a membrane, including transport as a result of signal sequences in its amino acid sequence when it is expressed in a suitable host cell. "Secreted" proteins include without limitation

proteins secreted wholly (*e.g.*, soluble proteins) or partially (*e.g.*, receptors) from the cell in which they are expressed. "Secreted" proteins also include without limitation proteins that are transported across the membrane of the endoplasmic reticulum. "Secreted" proteins are also intended to include proteins containing non-typical signal sequences (e.g. Interleukin-1 Beta, see Krasney, P.A. and Young, P.R. (1992) Cytokine 4(2):134 -143) and factors released from damaged cells (e.g. Interleukin-1 Receptor Antagonist, see Arend, W.P. et. al. (1998) Annu. Rev. Immunol. 16:27-55)

Where desired, an expression vector may be designed to contain a "signal or leader sequence" which will direct the polypeptide through the membrane of a cell. Such a sequence may be naturally present on the polypeptides of the present invention or provided from heterologous protein sources by recombinant DNA techniques.

The term "stringent" is used to refer to conditions that are commonly understood in the art as stringent. Stringent conditions can include highly stringent conditions (*i.e.*, hybridization to filter-bound DNA in 0.5 M $NaHPO_4$, 7% sodium dodecyl sulfate (SDS), 1 mM EDTA at 65°C, and washing in 0.1X SSC/0.1% SDS at 68°C), and moderately stringent conditions (*i.e.*, washing in 0.2X SSC/0.1% SDS at 42°C). Other exemplary hybridization conditions are described herein in the examples.

In instances of hybridization of deoxyoligonucleotides, additional exemplary stringent hybridization conditions include washing in 6X SSC/0.05% sodium pyrophosphate at 37°C (for 14-base oligonucleotides), 48°C (for 17-base oligos), 55°C (for 20-base oligonucleotides), and 60°C (for 23-base oligonucleotides).

As used herein, "substantially equivalent" can refer both to nucleotide and amino acid sequences, for example a mutant sequence, that varies from a reference sequence by one or more substitutions, deletions, or additions, the net effect of which does not result in an adverse functional dissimilarity between the reference and subject sequences. Typically, such a substantially equivalent sequence varies from one of those listed herein by no more than about 35% (*i.e.*, the number of individual residue substitutions, additions, and/or deletions in a substantially equivalent sequence, as compared to the corresponding reference sequence, divided by the total number of residues in the substantially equivalent sequence is about 0.35 or less). Such a sequence is said to have 65% sequence identity to the listed sequence. In one embodiment, a substantially

equivalent, *e.g.*, mutant, sequence of the invention varies from a listed sequence by no more than 30% (70% sequence identity); in a variation of this embodiment, by no more than 25% (75% sequence identity); and in a further variation of this embodiment, by no more than 20% (80% sequence identity) and in a further variation of this embodiment, by no more than 10% (90% sequence identity) and in a further variation of this embodiment, by no more that 5% (95% sequence identity). Substantially equivalent, *e.g.*, mutant, amino acid sequences according to the invention preferably have at least 80% sequence identity with a listed amino acid sequence, more preferably at least 90% sequence identity. Substantially equivalent nucleotide sequences of the invention can have lower percent sequence identities, taking into account, for example, the redundancy or degeneracy of the genetic code. Preferably, nucleotide sequence has at least about 65% identity, more preferably at least about 75% identity, and most preferably at least about 95% identity. For the purposes of the present invention, sequences having substantially equivalent biological activity and substantially equivalent expression characteristics are considered substantially equivalent. For the purposes of determining equivalence, truncation of the mature sequence (*e.g.*, via a mutation which creates a spurious stop codon) should be disregarded. Sequence identity may be determined, *e.g.*, using the Jotun Hein method (Hein, J. (1990) Methods Enzymol. 183:626-645). Identity between sequences can also be determined by other methods known in the art, e.g. by varying hybridization conditions.

The term "totipotent" refers to the capability of a cell to differentiate into all of the cell types of an adult organism.

The term "transformation" means introducing DNA into a suitable host cell so that the DNA is replicable, either as an extrachromosomal element, or by chromosomal integration. The term "transfection" refers to the taking up of an expression vector by a suitable host cell, whether or not any coding sequences are in fact expressed. The term "infection" refers to the introduction of nucleic acids into a suitable host cell by use of a virus or viral vector.

As used herein, an "uptake modulating fragment," UMF, means a series of nucleotides which mediate the uptake of a linked DNA fragment into a cell. UMFs can be readily identified using known UMFs as a target sequence or target motif with the

computer-based systems described below. The presence and activity of a UMF can be confirmed by attaching the suspected UMF to a marker sequence. The resulting nucleic acid molecule is then incubated with an appropriate host under appropriate conditions and the uptake of the marker sequence is determined. As described above, a UMF will increase the frequency of uptake of a linked marker sequence.

Each of the above terms is meant to encompass all that is described for each, unless the context dictates otherwise.

5.2 NUCLEIC ACIDS AND PEPTIDES OF THE INVENTION

Sequences of the nucleic acids and peptides of the present invention are set forth in the Sequence Listing. Table 1 relates the SEQ ID's listed herein to their identification in parent applications from which this application claims priority, and to corresponding SEQ ID NOs in the accompanying Sequence Listing.

TABLE 1

Gene Family	Serial Number	Date Filed	SEQ ID NO. in Parent Applications	SEQ ID NO: in Current Application	SEQ ID NO: In Sequence Listing
748	09/205,070	Dec. 3, 1998	SEQ ID NO: 1-45,207	748 SEQ ID NO: 1-45,207	1-45,207
	09/340,623	Jun. 28, 1999	SEQ ID NO: 1-45,207	748 SEQ ID NO: 1-45,207	1-45,207
	09/898,888	Jul. 3, 2001	SEQ ID NO: 1-45,207	748 SEQ ID NO: 1-45,207	1-45,207
752	09/205,155	Dec. 3, 1998	SEQ ID NO: 1-13203	752 SEQ ID NO: 1-13203	45,208-58,410
	09/359,922	Jul. 22, 1999	SEQ ID NO: 1-13203	752 SEQ ID NO: 1-13203	45,208-58,410
	09/919,002	Jul. 30, 2001	SEQ ID NO: 1-13203	752 SEQ ID NO: 1-13203	45,208-58,410
778	09/347,127	Jul. 2, 1999	SEQ ID NO: 1-105	778 SEQ ID NO: 1-105	58,411-58,515
	09/905,059	Jul. 12, 2001	SEQ ID NO: 1-105	778 SEQ ID NO: 1-105	58,411-58,515
779	09/457,877	Dec. 8, 1999	SEQ ID NO: 1-128	779 SEQ ID NO: 1-128	58,516-58,643
	09/952,981	Sep. 14, 2001	SEQ ID NO: 1-128	779 SEQ ID NO: 1-128	58,516-58,643
782	09/471,275	Dec. 23, 1999	SEQ ID NO: 1-10,451	782 SEQ ID NO: 1-10,451	58,644-69,094
784	09/488,725	Jan. 21, 2000	SEQ ID NO: 1-10289	784 SEQ ID NO: 1-10289	69,095-79,383
	09/552,317	Apr. 25, 2000	SEQ ID NO: 1-10289	784 SEQ ID NO: 1-10289	69,095-79,383
785	09/491,404	Jan. 25, 2000	SEQ ID NO: 1-3796	785 SEQ ID NO: 1-3796	79,384-83,179

	09/922,279	Aug. 3, 2001	SEQ ID NO: 1-3796	785 SEQ ID NO: 1-3796	79,384-83,179
787	09/496,914	Feb. 23, 2000	SEQ ID NO: 1-10,410	787 SEQ ID NO: 1-10,410	83,180-93,589
	09/560,875	Apr. 27, 2000	SEQ ID NO: 1-10,410	787 SEQ ID NO: 1-10,410	83,180-93,589
788	09/515,126	Feb. 28, 2000	SEQ ID NO: 1-14074	788 SEQ ID NO: 1-14074	93,590-107,663
	09/577,409	May. 18, 2000	SEQ ID NO: 1-14074	788 SEQ ID NO: 1-14074	93,590-107,663
789	09/519,705	Mar. 7, 2000	SEQ ID NO: 1-6391	789 SEQ ID NO: 1-6391	107,664-114,054
	09/574,454	May. 19, 2000	SEQ ID NO: 1-6391	789 SEQ ID NO: 1-6391	107,664-114,054
790	09/540,217	Mar. 31, 2000	SEQ ID NO: 1-30533	790 SEQ ID NO: 1-30533	114,055-144,607
	09/649,167	Aug. 23, 2000	SEQ ID NO: 1-30533	790 SEQ ID NO: 1-30533	114,055-144,607
791	09/552,929	Apr. 18, 2000	SEQ ID NO: 1-5822	791 SEQ ID NO: 1-5822	144,608-150,429
	09/770,160	Jan. 26, 2001	SEQ ID NO: 1-5822	791 SEQ ID NO: 1-5822	144,608-150,429
792	09/577,408	May. 18, 2000	SEQ ID NO: 1-8502	792 SEQ ID NO: 1-8502	150,430-158,931

The isolated polynucleotides of the invention include a polynucleotide comprising the nucleotide sequences of 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502; a polynucleotide encoding any one of the peptide sequences of 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502; and a polynucleotide comprising the nucleotide sequence encoding the mature protein coding sequence of the polypeptides of any one of 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502. The polynucleotides of the present invention also include, but are not limited to, a

polynucleotide that hybridizes under stringent conditions to (a) the complement of any of the nucleotides sequences of 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502 ; (b) nucleotide sequences encoding any one of the amino acid sequences set forth in the Sequence Listing; (c) a polynucleotide which is an allelic variant of any polynucleotide recited above; (d) a polynucleotide which encodes a species homolog of any of the proteins recited above; or (e) a polynucleotide that encodes a polypeptide comprising a specific domain or truncation of the polypeptides encoded by 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502. Domains of interest may depend on the nature of the encoded polypeptide; *e.g.*, domains in receptor-like polypeptides include ligand-binding, extracellular, transmembrane, or cytoplasmic domains, or combinations thereof; domains in immunoglobulin-like proteins include the variable immunoglobulin-like domains; domains in enzyme-like polypeptides include catalytic and substrate binding domains; and domains in ligand polypeptides include receptor-binding domains.

The polynucleotides of the invention include naturally occurring or wholly or partially synthetic DNA, *e.g.*, cDNA and genomic DNA, and RNA, *e.g.*, mRNA. The polynucleotides may include all of the coding region of the cDNA or may represent a portion of the coding region of the cDNA.

The present invention also provides genes corresponding to the cDNA sequences disclosed herein. The corresponding genes can be isolated in accordance with known methods using the sequence information disclosed herein. Such methods include the preparation of probes or primers from the disclosed sequence information for identification and/or amplification of genes in appropriate genomic libraries or other sources of genomic materials. Further 5' and 3' sequence can be obtained using methods known in the art. For example, full length cDNA or genomic DNA that corresponds to any of the polynucleotides

of 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502 can be obtained by screening appropriate cDNA or genomic DNA libraries under suitable hybridization conditions using any of the polynucleotides of 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502 or a portion thereof as a probe. Alternatively, the polynucleotides of 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502 may be used as the basis for suitable primer(s) that allow identification and/or amplification of genes in appropriate genomic DNA or cDNA libraries.

The nucleic acid sequences of the invention can be assembled from ESTs and sequences (including cDNA and genomic sequences) obtained from one or more public databases, such as dbEST, gbpri, and UniGene. The EST sequences can provide identifying sequence information, representative fragment or segment information, or novel segment information for the full-length gene.

The polynucleotides of the invention also provide polynucleotides including nucleotide sequences that are substantially equivalent to the polynucleotides recited above. Polynucleotides according to the invention can have, *e.g.*, at least about 65%, at least about 70%, at least about 75%, at least about 80%, more typically at least about 90%, and even more typically at least about 95%, sequence identity to a polynucleotide recited above.

Included within the scope of the nucleic acid sequences of the invention are nucleic acid sequence fragments that hybridize under stringent conditions to any of the nucleotide sequences of 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ

ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502, or complements thereof, which fragment is greater than about 5 nucleotides, preferably 7 nucleotides, more preferably greater than 9 nucleotides and most preferably greater than 17 nucleotides. Fragments of, e.g. 15, 17, or 20 nucleotides or more that are selective for (i.e. specifically hybridize to any one of the polynucleotides of the invention) are contemplated. Probes capable of specifically hybridizing to a polynucleotide can differentiate polynucleotide sequences of the invention from other polynucleotide sequences in the same family of genes or can differentiate human genes from genes of other species, and are preferably based on unique nucleotide sequences.

The sequences falling within the scope of the present invention are not limited to these specific sequences, but also include allelic and species variations thereof. Allelic and species variations can be routinely determined by comparing the sequence provided 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502, a representative fragment thereof, or a nucleotide sequence at least 90% identical, preferably 95% identical, to 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502 with a sequence from another isolate of the same species. Furthermore, to accommodate codon variability, the invention includes nucleic acid molecules coding for the same amino acid sequences as do the specific ORFs disclosed herein. In other words, in the coding region of an ORF, substitution of one codon for another codon that encodes the same amino acid is expressly contemplated.

The nearest neighbor or homology result for the nucleic acids of the present invention, including 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID

NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502, can be obtained by searching a database using an algorithm or a program. Preferably, a BLAST which stands for Basic Local Alignment Search Tool is used to search for local sequence alignments (Altshul, S.F. J Mol. Evol. 36 290-300 (1993) and Altschul S.F. et. al., J. Mol. Biol. 21:403-410 (1990)). Alternatively a FASTA version 3 search against Genpept, using Fastxy algorithm.

Species homologs (or orthologs) of the disclosed polynucleotides and proteins are also provided by the present invention. Species homologs may be isolated and identified by making suitable probes or primers from the sequences provided herein and screening a suitable nucleic acid source from the desired species.

The invention also encompasses allelic variants of the disclosed polynucleotides or proteins; that is, naturally-occurring alternative forms of the isolated polynucleotide which also encode proteins which are identical, homologous or related to that encoded by the polynucleotides.

The nucleic acid sequences of the invention are further directed to sequences which encode variants of the described nucleic acids. These amino acid sequence variants may be prepared by methods known in the art by introducing appropriate nucleotide changes into a native or variant polynucleotide. There are two variables in the construction of amino acid sequence variants: the location of the mutation and the nature of the mutation. Nucleic acids encoding the amino acid sequence variants are preferably constructed by mutating the polynucleotide to encode an amino acid sequence that does not occur in nature. These nucleic acid alterations can be made at sites that differ in the nucleic acids from different species (variable positions) or in highly conserved regions (constant regions). Sites at such locations will typically be modified in series, *e.g.*, by substituting first with conservative choices (*e.g.*, hydrophobic amino acid to a different hydrophobic amino acid) and then with more distant choices (*e.g.*, hydrophobic amino acid to a charged amino acid), and then deletions or insertions may be made at the target site. Amino acid sequence deletions generally range from about 1 to 30 residues, preferably about 1 to 10 residues, and are typically contiguous. Amino acid insertions

include amino- and/or carboxyl-terminal fusions ranging in length from one to one hundred or more residues, as well as intrasequence insertions of single or multiple amino acid residues. Intrasequence insertions may range generally from about 1 to 10 amino residues, preferably from 1 to 5 residues. Examples of terminal insertions include the heterologous signal sequences necessary for secretion or for intracellular targeting in different host cells and sequences such as FLAG or poly-histidine sequences useful for purifying the expressed protein.

In a preferred method, polynucleotides encoding the novel amino acid sequences are changed via site-directed mutagenesis. This method uses oligonucleotide sequences to alter a polynucleotide to encode the desired amino acid variant, as well as sufficient adjacent nucleotides on both sides of the changed amino acid to form a stable duplex on either side of the site of being changed. In general, the techniques of site-directed mutagenesis are well known to those of skill in the art and this technique is exemplified by publications such as, Edelman *et al., DNA* 2:183 (1983). A versatile and efficient method for producing site-specific changes in a polynucleotide sequence was published by Zoller and Smith, *Nucleic Acids Res.* 10:6487-6500 (1982). PCR may also be used to create amino acid sequence variants of the novel nucleic acids. When small amounts of template DNA are used as starting material, primer(s) that differs slightly in sequence from the corresponding region in the template DNA can generate the desired amino acid variant. PCR amplification results in a population of product DNA fragments that differ from the polynucleotide template encoding the polypeptide at the position specified by the primer. The product DNA fragments replace the corresponding region in the plasmid and this gives a polynucleotide encoding the desired amino acid variant.

A further technique for generating amino acid variants is the cassette mutagenesis technique described in Wells *et al., Gene* 34:315 (1985); and other mutagenesis techniques well known in the art, such as, for example, the techniques in Sambrook *et al.*, supra, and *Current Protocols in Molecular Biology*, Ausubel et. al., Due to the inherent degeneracy of the genetic code, other DNA sequences which encode substantially the same or a functionally equivalent amino acid sequence may be used in the practice of the invention for the cloning and expression of these novel nucleic acids. Such DNA

sequences include those which are capable of hybridizing to the appropriate novel nucleic acid sequence under stringent conditions.

Polynucleotides encoding preferred polypeptide truncations of the invention can be used to generate polynucleotides encoding chimeric or fusion proteins comprising one or more domains of the invention and heterologous protein sequences.

The polynucleotides of the invention additionally include the complement of any of the polynucleotides recited above. The polynucleotide can be DNA (genomic, cDNA, amplified, or synthetic) or RNA. Methods and algorithms for obtaining such polynucleotides are well known to those of skill in the art and can include, for example, methods for determining hybridization conditions that can routinely isolate polynucleotides of the desired sequence identities.

In accordance with the invention, polynucleotide sequences comprising the mature protein coding sequences corresponding to any one of 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502, or functional equivalents thereof, may be used to generate recombinant DNA molecules that direct the expression of that nucleic acid, or a functional equivalent thereof, in appropriate host cells. Also included are the cDNA inserts of any of the clones identified herein.

A polynucleotide according to the invention can be joined to any of a variety of other nucleotide sequences by well-established recombinant DNA techniques (see Sambrook J et. al., (1989) Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory, NY). Useful nucleotide sequences for joining to polynucleotides include an assortment of vectors, *e.g.*, plasmids, cosmids, lambda phage derivatives, phagemids, and the like, that are well known in the art. Accordingly, the invention also provides a vector including a polynucleotide of the invention and a host cell containing the polynucleotide. In general, the vector contains an origin of replication functional in at least one organism, convenient restriction endonuclease sites, and a selectable marker for the host cell. Vectors according to the invention include expression vectors, replication vectors, probe generation vectors, and sequencing vectors. A host cell according to the invention can be

a prokaryotic or eukaryotic cell and can be a unicellular organism or part of a multicellular organism.

The present invention further provides recombinant constructs comprising a nucleic acid having any of the nucleotide sequences of 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502 or a fragment thereof or any other polynucleotides of the invention. In one embodiment, the recombinant constructs of the present invention comprise a vector, such as a plasmid or viral vector, into which a nucleic acid having any of the nucleotide sequences of 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502 or a fragment thereof is inserted, in a forward or reverse orientation. In the case of a vector comprising one of the ORFs of the present invention, the vector may further comprise regulatory sequences, including for example, a promoter, operably linked to the ORF. Large numbers of suitable vectors and promoters are known to those of skill in the art and are commercially available for generating the recombinant constructs of the present invention. The following vectors are provided by way of example. Bacterial: pBs, phagescript, PsiX174, pBluescript SK, pBs KS, pNH8a, pNH16a, pNH18a, pNH46a (Stratagene); pTrc99A, pKK223-3, pKK233-3, pDR540, pRIT5 (Pharmacia). Eukaryotic: pWLneo, pSV2cat, pOG44, PXTI, pSG (Stratagene) pSVK3, pBPV, pMSG, pSVL (Pharmacia).

The isolated polynucleotide of the invention may be operably linked to an expression control sequence such as the pMT2 or pED expression vectors disclosed in Kaufman *et al., Nucleic Acids Res.* 19, 4485-4490 (1991), in order to produce the protein recombinantly. Many suitable expression control sequences are known in the art. General methods of expressing recombinant proteins are also known and are exemplified in R. Kaufman, *Methods in Enzymology* 185, 537-566 (1990). As defined herein

"operably linked" means that the isolated polynucleotide of the invention and an expression control sequence are situated within a vector or cell in such a way that the protein is expressed by a host cell which has been transformed (transfected) with the ligated polynucleotide/expression control sequence.

Promoter regions can be selected from any desired gene using CAT (chloramphenicol transferase) vectors or other vectors with selectable markers. Two appropriate vectors are pKK232-8 and pCM7. Particular named bacterial promoters include lacI, lacZ, T3, T7, gpt, lambda PR, and trc. Eukaryotic promoters include CMV immediate early, HSV thymidine kinase, early and late SV40, LTRs from retrovirus, and mouse metallothionein-I. Selection of the appropriate vector and promoter is well within the level of ordinary skill in the art. Generally, recombinant expression vectors will include origins of replication and selectable markers permitting transformation of the host cell, *e.g.*, the ampicillin resistance gene of *E. coli* and *S. cerevisiae* TRP1 gene, and a promoter derived from a highly-expressed gene to direct transcription of a downstream structural sequence. Such promoters can be derived from operons encoding glycolytic enzymes such as 3-phosphoglycerate kinase (PGK), a-factor, acid phosphatase, or heat shock proteins, among others. The heterologous structural sequence is assembled in appropriate phase with translation initiation and termination sequences, and preferably, a leader sequence capable of directing secretion of translated protein into the periplasmic space or extracellular medium. Optionally, the heterologous sequence can encode a fusion protein including an amino terminal identification peptide imparting desired characteristics, *e.g.*, stabilization or simplified purification of expressed recombinant product. Useful expression vectors for bacterial use are constructed by inserting a structural DNA sequence encoding a desired protein together with suitable translation initiation and termination signals in operable reading phase with a functional promoter. The vector will comprise one or more phenotypic selectable markers and an origin of replication to ensure maintenance of the vector and to, if desirable, provide amplification within the host. Suitable prokaryotic hosts for transformation include *E. coli*, *Bacillus subtilis*, *Salmonella typhimurium* and various species within the genera *Pseudomonas*, *Streptomyces*, and *Staphylococcus*, although others may also be employed as a matter of choice.

As a representative but non-limiting example, useful expression vectors for bacterial use can comprise a selectable marker and bacterial origin of replication derived from commercially available plasmids comprising genetic elements of the well known cloning vector pBR322 (ATCC 37017). Such commercial vectors include, for example, pKK223-3 (Pharmacia Fine Chemicals, Uppsala, Sweden) and GEM 1 (Promega Biotech, Madison, WI, USA). These pBR322 "backbone" sections are combined with an appropriate promoter and the structural sequence to be expressed. Following transformation of a suitable host strain and growth of the host strain to an appropriate cell density, the selected promoter is induced or derepressed by appropriate means (*e.g.*, temperature shift or chemical induction) and cells are cultured for an additional period. Cells are typically harvested by centrifugation, disrupted by physical or chemical means, and the resulting crude extract retained for further purification.

Polynucleotides of the invention can also be used to induce immune responses. For example, as described in Fan *et al., Nat. Biotech.* 17:870-872 (1999), incorporated herein by reference, nucleic acid sequences encoding a polypeptide may be used to generate antibodies against the encoded polypeptide following topical administration of naked plasmid DNA or following injection, and preferably intramuscular injection of the DNA. The nucleic acid sequences are preferably inserted in a recombinant expression vector and may be in the form of naked DNA.

5.3 ANTISENSE

Another aspect of the invention pertains to isolated antisense nucleic acid molecules that are hybridizable to or complementary to the nucleic acid molecule comprising the nucleotide sequence of 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502, or fragments, analogs or derivatives thereof. An "antisense" nucleic acid comprises a nucleotide sequence that is complementary to a "sense" nucleic acid encoding a protein, *e.g.*, complementary to the coding strand of a double-stranded cDNA molecule or complementary to an mRNA

sequence. In specific aspects, antisense nucleic acid molecules are provided that comprise a sequence complementary to at least about 10, 25, 50, 100, 250 or 500 nucleotides or an entire coding strand, or to only a portion thereof. Nucleic acid molecules encoding fragments, homologs, derivatives and analogs of a protein of any of 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502 or antisense nucleic acids complementary to a nucleic acid sequence of 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502 are additionally provided.

In one embodiment, an antisense nucleic acid molecule is antisense to a "coding region" of the coding strand of a nucleotide sequence of the invention. The term "coding region" refers to the region of the nucleotide sequence comprising codons which are translated into amino acid residues. In another embodiment, the antisense nucleic acid molecule is antisense to a "noncoding region" of the coding strand of a nucleotide sequence of the invention. The term "noncoding region" refers to 5' and 3' sequences which flank the coding region that are not translated into amino acids (*i.e.*, also referred to as 5' and 3' untranslated regions).

Given the coding strand sequences encoding a nucleic acid disclosed herein (*e.g.*, 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502, antisense nucleic acids of the invention can be designed according to the rules of Watson and Crick or Hoogsteen base pairing. The antisense nucleic acid molecule can be complementary to the entire coding region of a mRNA, but more preferably is an oligonucleotide that is antisense to only a portion of the coding or noncoding region of a

mRNA. For example, the antisense oligonucleotide can be complementary to the region surrounding the translation start site of a mRNA. An antisense oligonucleotide can be, for example, about 5, 10, 15, 20, 25, 30, 35, 40, 45 or 50 nucleotides in length. An antisense nucleic acid of the invention can be constructed using chemical synthesis or enzymatic ligation reactions using procedures known in the art. For example, an antisense nucleic acid (*e.g.*, an antisense oligonucleotide) can be chemically synthesized using naturally occurring nucleotides or variously modified nucleotides designed to increase the biological stability of the molecules or to increase the physical stability of the duplex formed between the antisense and sense nucleic acids, *e.g.*, phosphorothioate derivatives and acridine substituted nucleotides can be used.

Examples of modified nucleotides that can be used to generate the antisense nucleic acid include: 5-fluorouracil, 5-bromouracil, 5-chlorouracil, 5-iodouracil, hypoxanthine, xanthine, 4-acetylcytosine, 5-(carboxyhydroxylmethyl) uracil, 5-carboxymethylaminomethyl-2-thiouridine, 5-carboxymethylaminomethyluracil, dihydrouracil, beta-D-galactosylqueosine, inosine, n6-isopentenyladenine, 1-methylguanine, 1-methylinosine, 2,2-dimethylguanine, 2-methyladenine, 2-methylguanine, 3-methylcytosine, 5-methylcytosine, n6-adenine, 7-methylguanine, 5-methylaminomethyluracil, 5-methoxyaminomethyl-2-thiouracil, beta-D-mannosylqueosine, 5'-methoxycarboxymethyluracil, 5-methoxyuracil, 2-methylthio-n6-isopentenyladenine, uracil-5-oxyacetic acid (v), wybutoxosine, pseudouracil, queosine, 2-thiocytosine, 5-methyl-2-thiouracil, 2-thiouracil, 4-thiouracil, 5-methyluracil, uracil-5-oxyacetic acid methylester, uracil-5-oxyacetic acid (v), 5-methyl-2-thiouracil, 3-(3-amino-3-N-2-carboxypropyl) uracil, (acp3)w, and 2,6-diaminopurine. Alternatively, the antisense nucleic acid can be produced biologically using an expression vector into which a nucleic acid has been subcloned in an antisense orientation (*i.e.*, RNA transcribed from the inserted nucleic acid will be of an antisense orientation to a target nucleic acid of interest, described further in the following subsection).

The antisense nucleic acid molecules of the invention are typically administered to a subject or generated *in situ* such that they hybridize with or bind to cellular mRNA and/or genomic DNA encoding a protein according to the invention to thereby inhibit

expression of the protein, *e.g.*, by inhibiting transcription and/or translation. The hybridization can be by conventional nucleotide complementarity to form a stable duplex, or, for example, in the case of an antisense nucleic acid molecule that binds to DNA duplexes, through specific interactions in the major groove of the double helix. An example of a route of administration of antisense nucleic acid molecules of the invention includes direct injection at a tissue site. Alternatively, antisense nucleic acid molecules can be modified to target selected cells and then administered systemically. For example, for systemic administration, antisense molecules can be modified such that they specifically bind to receptors or antigens expressed on a selected cell surface, *e.g.*, by linking the antisense nucleic acid molecules to peptides or antibodies that bind to cell surface receptors or antigens. The antisense nucleic acid molecules can also be delivered to cells using the vectors described herein. To achieve sufficient intracellular concentrations of antisense molecules, vector constructs in which the antisense nucleic acid molecule is placed under the control of a strong pol II or pol III promoter are preferred.

In yet another embodiment, the antisense nucleic acid molecule of the invention is an α-anomeric nucleic acid molecule. An α-anomeric nucleic acid molecule forms specific double-stranded hybrids with complementary RNA in which, contrary to the usual β-units, the strands run parallel to each other (Gaultier *et al.*, (1987) *Nucleic Acids Res* 15: 6625-6641). The antisense nucleic acid molecule can also comprise a 2'-o-methylribonucleotide (Inoue *et al.*, (1987) *Nucleic Acids Res* 15: 6131-6148) or a chimeric RNA -DNA analogue (Inoue *et al.*, (1987) *FEBS Lett* 215: 327-330).

5.4 RIBOZYMES AND PNA MOIETIES

In still another embodiment, an antisense nucleic acid of the invention is a ribozyme. Ribozymes are catalytic RNA molecules with ribonuclease activity that are capable of cleaving a single-stranded nucleic acid, such as an mRNA, to which they have a complementary region. Thus, ribozymes (*e.g.*, hammerhead ribozymes (described in Haselhoff and Gerlach (1988) *Nature* 334:585-591)) can be used to catalytically cleave mRNA transcripts to thereby inhibit translation of the mRNA. A ribozyme having specificity for a nucleic acid of the invention can be designed based upon the nucleotide

sequence of a DNA disclosed herein (*i.e.*, 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502). For example, a derivative of a Tetrahymena L-19 IVS RNA can be constructed in which the nucleotide sequence of the active site is complementary to the nucleotide sequence to be cleaved in a SECX-encoding mRNA. See, *e.g.*, Cech *et al.*, U.S. Pat. No. 4,987,071; and Cech *et al.*, U.S. Pat. No. 5,116,742. Alternatively, SECX mRNA can be used to select a catalytic RNA having a specific ribonuclease activity from a pool of RNA molecules. See, *e.g.*, Bartel *et al.*, (1993) *Science* 261:1411-1418.

Alternatively, gene expression can be inhibited by targeting nucleotide sequences complementary to the regulatory region (*e.g.*, promoter and/or enhancers) to form triple helical structures that prevent transcription of the gene in target cells. See generally, Helene. (1991) *Anticancer Drug Des.* 6: 569-84; Helene. *et al.*, (1992) *Ann. N.Y. Acad. Sci.* 660:27-36; and Maher (1992) *Bioassays* 14: 807-15.

In various embodiments, the nucleic acids of the invention can be modified at the base moiety, sugar moiety or phosphate backbone to improve, *e.g.*, the stability, hybridization, or solubility of the molecule. For example, the deoxyribose phosphate backbone of the nucleic acids can be modified to generate peptide nucleic acids (see Hyrup *et al.*, (1996) *Bioorg Med Chem* 4: 5-23). As used herein, the terms "peptide nucleic acids" or "PNAs" refer to nucleic acid mimics, *e.g.*, DNA mimics, in which the deoxyribose phosphate backbone is replaced by a pseudopeptide backbone and only the four natural nucleobases are retained. The neutral backbone of PNAs has been shown to allow for specific hybridization to DNA and RNA under conditions of low ionic strength. The synthesis of PNA oligomers can be performed using standard solid phase peptide synthesis protocols as described in Hyrup *et al.*, (1996) above; Perry-O'Keefe *et al.*, (1996) *PNAS* 93: 14670-675.

PNAs of the invention can be used in therapeutic and diagnostic applications. For example, PNAs can be used as antisense or antigene agents for sequence-specific modulation of gene expression by, *e.g.*, inducing transcription or translation arrest or

inhibiting replication. PNAs of the invention can also be used, *e.g.*, in the analysis of single base pair mutations in a gene by, *e.g.*, PNA directed PCR clamping; as artificial restriction enzymes when used in combination with other enzymes, *e.g.*, S1 nucleases (Hyrup B. (1996) above); or as probes or primers for DNA sequence and hybridization (Hyrup *et al.*, (1996), above; Perry-O'Keefe (1996), above).

In another embodiment, PNAs of the invention can be modified, *e.g.*, to enhance their stability or cellular uptake, by attaching lipophilic or other helper groups to PNA, by the formation of PNA-DNA chimeras, or by the use of liposomes or other techniques of drug delivery known in the art. For example, PNA-DNA chimeras can be generated that may combine the advantageous properties of PNA and DNA. Such chimeras allow DNA recognition enzymes, *e.g.*, RNase H and DNA polymerases, to interact with the DNA portion while the PNA portion would provide high binding affinity and specificity. PNA-DNA chimeras can be linked using linkers of appropriate lengths selected in terms of base stacking, number of bonds between the nucleobases, and orientation (Hyrup (1996) above). The synthesis of PNA-DNA chimeras can be performed as described in Hyrup (1996) above and Finn *et al.*, (1996) *Nucl Acids Res* 24: 3357-63. For example, a DNA chain can be synthesized on a solid support using standard phosphoramidite coupling chemistry, and modified nucleoside analogs, *e.g.*, 5'-(4-methoxytrityl)amino-5'-deoxy-thymidine phosphoramidite, can be used between the PNA and the 5' end of DNA (Mag *et al.*, (1989) *Nucl Acid Res* 17: 5973-88). PNA monomers are then coupled in a stepwise manner to produce a chimeric molecule with a 5' PNA segment and a 3' DNA segment (Finn *et al.*, (1996) above). Alternatively, chimeric molecules can be synthesized with a 5' DNA segment and a 3' PNA segment. See, Petersen *et al.*, (1975) *Bioorg Med Chem Lett* 5: 1119-11124.

In other embodiments, the oligonucleotide may include other appended groups such as peptides (*e.g.*, for targeting host cell receptors *in vivo*), or agents facilitating transport across the cell membrane (see, *e.g.*, Letsinger *et al.*,, 1989, *Proc. Natl. Acad. Sci. U.S.A.* 86:6553-6556; Lemaitre *et al.*,, 1987, Proc. *Natl. Acad. Sci.* 84:648-652; PCT Publication No. W088/09810) or the blood-brain barrier (see, *e.g.*, PCT Publication No. W089/10134). In addition, oligonucleotides can be modified with hybridization triggered cleavage agents (See, *e.g.*, Krol *et al.*,, 1988, *BioTechniques* 6:958-976) or intercalating

agents. (See, *e.g.*, Zon, 1988, *Pharm. Res.* 5: 539-549). To this end, the oligonucleotide may be conjugated to another molecule, *e.g.*, a peptide, a hybridization triggered cross-linking agent, a transport agent, a hybridization-triggered cleavage agent, etc.

5.5 HOSTS

The present invention further provides host cells genetically engineered to contain the polynucleotides of the invention. For example, such host cells may contain nucleic acids of the invention introduced into the host cell using known transformation, transfection or infection methods. The present invention still further provides host cells genetically engineered to express the polynucleotides of the invention, wherein such polynucleotides are in operative association with a regulatory sequence heterologous to the host cell which drives expression of the polynucleotides in the cell.

Knowledge of nucleic acid sequences allows for modification of cells to permit, or increase, expression of endogenous polypeptide. Cells can be modified (*e.g.*, by homologous recombination) to provide increased polypeptide expression by replacing, in whole or in part, the naturally occurring promoter with all or part of a heterologous promoter so that the cells express the polypeptide at higher levels. The heterologous promoter is inserted in such a manner that it is operatively linked to the encoding sequences. See, for example, PCT International Publication No. WO94/12650, PCT International Publication No. WO92/20808, and PCT International Publication No. WO91/09955. It is also contemplated that, in addition to heterologous promoter DNA, amplifiable marker DNA (*e.g.*, ada, dhfr, and the multifunctional CAD gene which encodes carbamyl phosphate synthase, aspartate transcarbamylase, and dihydroorotase) and/or intron DNA may be inserted along with the heterologous promoter DNA. If linked to the coding sequence, amplification of the marker DNA by standard selection methods results in co-amplification of the desired protein coding sequences in the cells.

The host cell can be a higher eukaryotic host cell, such as a mammalian cell, a lower eukaryotic host cell, such as a yeast cell, or the host cell can be a prokaryotic cell, such as a bacterial cell. Introduction of the recombinant construct into the host cell can be effected by calcium phosphate transfection, DEAE, dextran mediated transfection, or electroporation (Davis, L. *et al., Basic Methods in Molecular Biology* (1986)). The host

cells containing one of the polynucleotides of the invention, can be used in conventional manners to produce the gene product encoded by the isolated fragment (in the case of an ORF) or can be used to produce a heterologous protein under the control of the EMF.

Any host/vector system can be used to express one or more of the ORFs of the present invention. These include, but are not limited to, eukaryotic hosts such as HeLa cells, Cv-1 cell, COS cells, 293 cells, and Sf9 cells, as well as prokaryotic host such as *E. coli* and *B. subtilis*. The most preferred cells are those which do not normally express the particular polypeptide or protein or which expresses the polypeptide or protein at low natural level. Mature proteins can be expressed in mammalian cells, yeast, bacteria, or other cells under the control of appropriate promoters. Cell-free translation systems can also be employed to produce such proteins using RNAs derived from the DNA constructs of the present invention. Appropriate cloning and expression vectors for use with prokaryotic and eukaryotic hosts are described by Sambrook, *et al.*, in Molecular Cloning: A Laboratory Manual, Second Edition, Cold Spring Harbor, New York (1989), the disclosure of which is hereby incorporated by reference.

Various mammalian cell culture systems can also be employed to express recombinant protein. Examples of mammalian expression systems include the COS-7 lines of monkey kidney fibroblasts, described by Gluzman, Cell 23:175 (1981). Other cell lines capable of expressing a compatible vector are, for example, the C127, monkey COS cells, Chinese Hamster Ovary (CHO) cells, human kidney 293 cells, human epidermal A431 cells, human Colo205 cells, 3T3 cells, CV-1 cells, other transformed primate cell lines, normal diploid cells, cell strains derived from *in vitro* culture of primary tissue, primary explants, HeLa cells, mouse L cells, BHK, HL-60, U937, HaK or Jurkat cells. Mammalian expression vectors will comprise an origin of replication, a suitable promoter and also any necessary ribosome binding sites, polyadenylation site, splice donor and acceptor sites, transcriptional termination sequences, and 5′ flanking nontranscribed sequences. DNA sequences derived from the SV40 viral genome, for example, SV40 origin, early promoter, enhancer, splice, and polyadenylation sites may be used to provide the required nontranscribed genetic elements. Recombinant polypeptides and proteins produced in bacterial culture are usually isolated by initial extraction from cell pellets, followed by one or more salting-out, aqueous ion exchange or size exclusion

chromatography steps. Protein refolding steps can be used, as necessary, in completing configuration of the mature protein. Finally, high performance liquid chromatography (HPLC) can be employed for final purification steps. Microbial cells employed in expression of proteins can be disrupted by any convenient method, including freeze-thaw cycling, sonication, mechanical disruption, or use of cell lysing agents.

Alternatively, it may be possible to produce the protein in lower eukaryotes such as yeast or insects or in prokaryotes such as bacteria. Potentially suitable yeast strains include *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Kluyveromyces* strains, *Candida*, or any yeast strain capable of expressing heterologous proteins. Potentially suitable bacterial strains include *Escherichia coli, Bacillus subtilis, Salmonella typhimurium*, or any bacterial strain capable of expressing heterologous proteins. If the protein is made in yeast or bacteria, it may be necessary to modify the protein produced therein, for example by phosphorylation or glycosylation of the appropriate sites, in order to obtain the functional protein. Such covalent attachments may be accomplished using known chemical or enzymatic methods.

In another embodiment of the present invention, cells and tissues may be engineered to express an endogenous gene comprising the polynucleotides of the invention under the control of inducible regulatory elements, in which case the regulatory sequences of the endogenous gene may be replaced by homologous recombination. As described herein, gene targeting can be used to replace a gene's existing regulatory region with a regulatory sequence isolated from a different gene or a novel regulatory sequence synthesized by genetic engineering methods. Such regulatory sequences may be comprised of promoters, enhancers, scaffold-attachment regions, negative regulatory elements, transcriptional initiation sites, regulatory protein binding sites or combinations of said sequences. Alternatively, sequences which affect the structure or stability of the RNA or protein produced may be replaced, removed, added, or otherwise modified by targeting. These sequence include polyadenylation signals, mRNA stability elements, splice sites, leader sequences for enhancing or modifying transport or secretion properties of the protein, or other sequences which alter or improve the function or stability of protein or RNA molecules.

The targeting event may be a simple insertion of the regulatory sequence, placing the gene under the control of the new regulatory sequence, *e.g.*, inserting a new promoter or enhancer or both upstream of a gene. Alternatively, the targeting event may be a simple deletion of a regulatory element, such as the deletion of a tissue-specific negative regulatory element. Alternatively, the targeting event may replace an existing element; for example, a tissue-specific enhancer can be replaced by an enhancer that has broader or different cell-type specificity than the naturally occurring elements. Here, the naturally occurring sequences are deleted and new sequences are added. In all cases, the identification of the targeting event may be facilitated by the use of one or more selectable marker genes that are contiguous with the targeting DNA, allowing for the selection of cells in which the exogenous DNA has integrated into the host cell genome. The identification of the targeting event may also be facilitated by the use of one or more marker genes exhibiting the property of negative selection, such that the negatively selectable marker is linked to the exogenous DNA, but configured such that the negatively selectable marker flanks the targeting sequence, and such that a correct homologous recombination event with sequences in the host cell genome does not result in the stable integration of the negatively selectable marker. Markers useful for this purpose include the Herpes Simplex Virus thymidine kinase (TK) gene or the bacterial xanthine-guanine phosphoribosyl-transferase (gpt) gene.

The gene targeting or gene activation techniques which can be used in accordance with this aspect of the invention are more particularly described in U.S. Patent No. 5,272,071 to Chappel; U.S. Patent No. 5,578,461 to Sherwin et. al.,; International Application No. PCT/US92/09627 (WO93/09222) by Selden et. al.,; and International Application No. PCT/US90/06436 (WO91/06667) by Skoultchi *et al.*, each of which is incorporated by reference herein in its entirety.

5.6 POLYPEPTIDES OF THE INVENTION

The isolated polypeptides of the invention include, but are not limited to, a polypeptide comprising: the amino acid sequences set forth as any one of 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796,

787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502 or an amino acid sequence encoded by any one of the nucleotide sequences 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502 or the corresponding full length or mature protein. Polypeptides of the invention also include polypeptides preferably with biological or immunological activity that are encoded by: (a) a polynucleotide having any one of the nucleotide sequences set forth in 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502 or (b) polynucleotides encoding any one of the amino acid sequences set forth 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502 or (c) polynucleotides that hybridize to the complement of the polynucleotides of either (a) or (b) under stringent hybridization conditions. The invention also provides biologically active or immunologically active variants of any of the amino acid sequences set forth as 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502or the corresponding full length or mature protein; and "substantial equivalents" thereof (*e.g.*, with at least about 65%, at least about 70%, at least about 75%, at least about 80%, at least about 85%, at least about 90%, typically at least about 95%, more typically at least about 98%, or most typically at least about 99% amino acid identity) that retain biological activity. Polypeptides encoded by allelic variants may have a

similar, increased, or decreased activity compared to polypeptides comprising 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502.

Fragments of the proteins of the present invention which are capable of exhibiting biological activity are also encompassed by the present invention. Fragments of the protein may be in linear form or they may be cyclized using known methods, for example, as described in H. U. Saragovi, *et al.*, Bio/Technology 10, 773-778 (1992) and in R. S. McDowell, *et al.*, J. Amer. Chem. Soc. 114, 9245-9253 (1992), both of which are incorporated herein by reference. Such fragments may be fused to carrier molecules such as immunoglobulins for many purposes, including increasing the valency of protein binding sites.

The present invention also provides both full-length and mature forms (for example, without a signal sequence or precursor sequence) of the disclosed proteins. The protein coding sequence is identified in the sequence listing by translation of the disclosed nucleotide sequences. The mature form of such protein may be obtained by expression of a full-length polynucleotide in a suitable mammalian cell or other host cell. The sequence of the mature form of the protein is also determinable from the amino acid sequence of the full-length form. Where proteins of the present invention are membrane bound, soluble forms of the proteins are also provided. In such forms, part or all of the regions causing the proteins to be membrane bound are deleted so that the proteins are fully secreted from the cell in which they are expressed.

Protein compositions of the present invention may further comprise an acceptable carrier, such as a hydrophilic, *e.g.*, pharmaceutically acceptable, carrier.

The present invention further provides isolated polypeptides encoded by the nucleic acid fragments of the present invention or by degenerate variants of the nucleic acid fragments of the present invention. By "degenerate variant" is intended nucleotide fragments which differ from a nucleic acid fragment of the present invention (*e.g.*, an ORF) by nucleotide sequence but, due to the degeneracy of the genetic code, encode an

identical polypeptide sequence. Preferred nucleic acid fragments of the present invention are the ORFs that encode proteins.

A variety of methodologies known in the art can be utilized to obtain any one of the isolated polypeptides or proteins of the present invention. At the simplest level, the amino acid sequence can be synthesized using commercially available peptide synthesizers. The synthetically-constructed protein sequences, by virtue of sharing primary, secondary or tertiary structural and/or conformational characteristics with proteins may possess biological properties in common therewith, including protein activity. This technique is particularly useful in producing small peptides and fragments of larger polypeptides. Fragments are useful, for example, in generating antibodies against the native polypeptide. Thus, they may be employed as biologically active or immunological substitutes for natural, purified proteins in screening of therapeutic compounds and in immunological processes for the development of antibodies.

The polypeptides and proteins of the present invention can alternatively be purified from cells which have been altered to express the desired polypeptide or protein. As used herein, a cell is said to be altered to express a desired polypeptide or protein when the cell, through genetic manipulation, is made to produce a polypeptide or protein which it normally does not produce or which the cell normally produces at a lower level. One skilled in the art can readily adapt procedures for introducing and expressing either recombinant or synthetic sequences into eukaryotic or prokaryotic cells in order to generate a cell which produces one of the polypeptides or proteins of the present invention.

The invention also relates to methods for producing a polypeptide comprising growing a culture of host cells of the invention in a suitable culture medium, and purifying the protein from the cells or the culture in which the cells are grown. For example, the methods of the invention include a process for producing a polypeptide in which a host cell containing a suitable expression vector that includes a polynucleotide of the invention is cultured under conditions that allow expression of the encoded polypeptide. The polypeptide can be recovered from the culture, conveniently from the culture medium, or from a lysate prepared from the host cells and further purified.

Preferred embodiments include those in which the protein produced by such process is a full length or mature form of the protein.

In an alternative method, the polypeptide or protein is purified from bacterial cells which naturally produce the polypeptide or protein. One skilled in the art can readily follow known methods for isolating polypeptides and proteins in order to obtain one of the isolated polypeptides or proteins of the present invention. These include, but are not limited to, immunochromatography, HPLC, size-exclusion chromatography, ion-exchange chromatography, and immuno-affinity chromatography. See, *e.g.*, Scopes, *Protein Purification: Principles and Practice*, Springer-Verlag (1994); Sambrook, *et al.*, in Molecular Cloning: *A Laboratory Manual*; Ausubel *et al., Current Protocols in Molecular Biology*. Polypeptide fragments that retain biological/immunological activity include fragments comprising greater than about 100 amino acids, or greater than about 200 amino acids, and fragments that encode specific protein domains.

The purified polypeptides can be used in *in vitro* binding assays which are well known in the art to identify molecules which bind to the polypeptides. These molecules include but are not limited to, for *e.g.*, small molecules, molecules from combinatorial libraries, antibodies or other proteins. The molecules identified in the binding assay are then tested for antagonist or agonist activity in *in vivo* tissue culture or animal models that are well known in the art. In brief, the molecules are titrated into a plurality of cell cultures or animals and then tested for either cell/animal death or prolonged survival of the animal/cells.

In addition, the peptides of the invention or molecules capable of binding to the peptides may be complexed with toxins, *e.g.*, ricin or cholera, or with other compounds that are toxic to cells. The toxin-binding molecule complex is then targeted to a tumor or other cell by the specificity of the binding molecule for 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502.

The protein of the invention may also be expressed as a product of transgenic animals, *e.g.*, as a component of the milk of transgenic cows, goats, pigs, or sheep which

are characterized by somatic or germ cells containing a nucleotide sequence encoding the protein.

The proteins provided herein also include proteins characterized by amino acid sequences similar to those of purified proteins but into which modification are naturally provided or deliberately engineered. For example, modifications, in the peptide or DNA sequence, can be made by those skilled in the art using known techniques. Modifications of interest in the protein sequences may include the alteration, substitution, replacement, insertion or deletion of a selected amino acid residue in the coding sequence. For example, one or more of the cysteine residues may be deleted or replaced with another amino acid to alter the conformation of the molecule. Techniques for such alteration, substitution, replacement, insertion or deletion are well known to those skilled in the art (see, *e.g.*, U.S. Pat. No. 4,518,584). Preferably, such alteration, substitution, replacement, insertion or deletion retains the desired activity of the protein. Regions of the protein that are important for the protein function can be determined by various methods known in the art including the alanine-scanning method which involved systematic substitution of single or strings of amino acids with alanine, followed by testing the resulting alanine-containing variant for biological activity. This type of analysis determines the importance of the substituted amino acid(s) in biological activity. Regions of the protein that are important for protein function may be determined by the eMATRIX program.

Other fragments and derivatives of the sequences of proteins which would be expected to retain protein activity in whole or in part and are useful for screening or other immunological methodologies may also be easily made by those skilled in the art given the disclosures herein. Such modifications are encompassed by the present invention.

The protein may also be produced by operably linking the isolated polynucleotide of the invention to suitable control sequences in one or more insect expression vectors, and employing an insect expression system. Materials and methods for baculovirus/insect cell expression systems are commercially available in kit form from, *e.g.*, Invitrogen, San Diego, Calif., U.S.A. (the MaxBat™ kit), and such methods are well known in the art, as described in Summers and Smith, Texas Agricultural Experiment Station Bulletin No. 1555 (1987), incorporated herein by reference. As used herein, an

insect cell capable of expressing a polynucleotide of the present invention is "transformed."

The protein of the invention may be prepared by culturing transformed host cells under culture conditions suitable to express the recombinant protein. The resulting expressed protein may then be purified from such culture (*i.e.*, from culture medium or cell extracts) using known purification processes, such as gel filtration and ion exchange chromatography. The purification of the protein may also include an affinity column containing agents which will bind to the protein; one or more column steps over such affinity resins as concanavalin A-agarose, heparin-toyopearl™ (TOSOH Biosep LLC, Montgomeryville, PA) or Cibacrom blue 3GA Sepharose™ (Pharmacia, Piscataway, NJ); one or more steps involving hydrophobic interaction chromatography using such resins as phenyl ether, butyl ether, or propyl ether; or immunoaffinity chromatography.

Alternatively, the protein of the invention may also be expressed in a form which will facilitate purification. For example, it may be expressed as a fusion protein, such as those of maltose binding protein (MBP), glutathione-S-transferase (GST) or thioredoxin (TRX), or as a His tag. Kits for expression and purification of such fusion proteins are commercially available from New England BioLab (Beverly, Mass.), Pharmacia (Piscataway, N.J.) and Invitrogen, respectively. The protein can also be tagged with an epitope and subsequently purified by using a specific antibody directed to such epitope. One such epitope ("FLAG®") is commercially available from Kodak (New Haven, Conn.).

Finally, one or more reverse-phase high performance liquid chromatography (RP-HPLC) steps employing hydrophobic RP-HPLC media, *e.g.*, silica gel having pendant methyl or other aliphatic groups, can be employed to further purify the protein. Some or all of the foregoing purification steps, in various combinations, can also be employed to provide a substantially homogeneous isolated recombinant protein. The protein thus purified is substantially free of other mammalian proteins and is defined in accordance with the present invention as an "isolated protein."

The polypeptides of the invention include analogs (variants). This embraces fragments, as well as peptides in which one or more amino acids has been deleted, inserted, or substituted. Also, analogs of the polypeptides of the invention embrace

fusions of the polypeptides or modifications of the polypeptides of the invention, wherein the polypeptide or analog is fused to another moiety or moieties, *e.g.*, targeting moiety or another therapeutic agent. Such analogs may exhibit improved properties such as activity and/or stability. Examples of moieties which may be fused to the polypeptide or an analog include, for example, targeting moieties which provide for the delivery of polypeptide to pancreatic cells, *e.g.*, antibodies to pancreatic cells, antibodies to immune cells such as T-cells, monocytes, dendritic cells, granulocytes, etc., as well as receptor and ligands expressed on pancreatic or immune cells. Other moieties which may be fused to the polypeptide include therapeutic agents which are used for treatment, for example, immunosuppressive drugs such as cyclosporin, SK506, azathioprine, CD3 antibodies and steroids. Also, polypeptides may be fused to immune modulators, and other cytokines such as alpha or beta interferon.

5.6.1 DETERMINING POLYPEPTIDE AND POLYNUCLEOTIDE IDENTITY AND SIMILARITY

Preferred identity and/or similarity are designed to give the largest match between the sequences tested. Methods to determine identity and similarity are codified in computer programs including, but are not limited to, the GCG program package, including GAP (Devereux, J., *et al.*, Nucleic Acids Research 12(1):387 (1984); Genetics Computer Group, University of Wisconsin, Madison, WI), BLASTP, BLASTN, BLASTX, FASTA (Altschul, S.F. *et al.*, J. Molec. Biol. 215:403-410 (1990), PSI-BLAST (Altschul S.F. *et al.*, Nucleic Acids Res. vol. 25, pp. 3389-3402, herein incorporated by reference), eMatrix software (Wu *et al.*, J. Comp. Biol., Vol. 6, pp. 219-235 (1999), herein incorporated by reference), eMotif software (Nevill-Manning et al, ISMB-97, Vol. 4, pp. 202-209, herein incorporated by reference), pFam software (Sonnhammer *et al.*, Nucleic Acids Res., Vol. 26(1), pp. 320-322 (1998), herein incorporated by reference) and the Kyte-Doolittle hydrophobocity prediction algorithm (J. Mol Biol, 157, pp. 105-31 (1982), incorporated herein by reference). The BLAST programs are publicly available from the National Center for Biotechnology Information (NCBI) and other sources (BLAST Manual, Altschul, S., et. al., NCB NLM NIH Bethesda, MD 20894; Altschul, S., *et al.*, J. Mol. Biol. 215:403-410 (1990).

5.7 CHIMERIC AND FUSION PROTEINS

The invention also provides chimeric or fusion proteins. As used herein, a "chimeric protein" or "fusion protein" comprises a polypeptide of the invention operatively linked to another polypeptide. Within a fusion protein the polypeptide according to the invention can correspond to all or a portion of a protein according to the invention. In one embodiment, a fusion protein comprises at least one biologically active portion of a protein according to the invention. In another embodiment, a fusion protein comprises at least two biologically active portions of a protein according to the invention. Within the fusion protein, the term "operatively linked" is intended to indicate that the polypeptide according to the invention and the other polypeptide are fused in-frame to each other. The polypeptide can be fused to the N-terminus or C-terminus.

For example, in one embodiment a fusion protein comprises a polypeptide according to the invention operably linked to the extracellular domain of a second protein. In another embodiment, the fusion protein is a GST-fusion protein in which the polypeptide sequences of the invention are fused to the C-terminus of the GST (*i.e.*, glutathione S-transferase) sequences.

In another embodiment, the fusion protein is an immunoglobulin fusion protein in which the polypeptide sequences according to the invention comprises one or more domains are fused to sequences derived from a member of the immunoglobulin protein family. The immunoglobulin fusion proteins of the invention can be incorporated into pharmaceutical compositions and administered to a subject to inhibit an interaction between a ligand and a protein of the invention on the surface of a cell, to thereby suppress signal transduction *in vivo*. The immunoglobulin fusion proteins can be used to affect the bioavailability of a cognate ligand. Inhibition of the ligand/protein interaction may be useful therapeutically for both the treatment of proliferative and differentiative disorders, *e,g.*, cancer as well as modulating (*e.g.*, promoting or inhibiting) cell survival. Moreover, the immunoglobulin fusion proteins of the invention can be used as immunogens to produce antibodies in a subject, to purify ligands, and in screening assays to identify molecules that inhibit the interaction of a polypeptide of the invention with a ligand.

A chimeric or fusion protein of the invention can be produced by standard recombinant DNA techniques. For example, DNA fragments coding for the different polypeptide sequences are ligated together in-frame in accordance with conventional techniques, *e.g.*, by employing blunt-ended or stagger-ended termini for ligation, restriction enzyme digestion to provide for appropriate termini, filling-in of cohesive ends as appropriate, alkaline phosphatase treatment to avoid undesirable joining, and enzymatic ligation. In another embodiment, the fusion gene can be synthesized by conventional techniques including automated DNA synthesizers. Alternatively, PCR amplification of gene fragments can be carried out using anchor primers that give rise to complementary overhangs between two consecutive gene fragments that can subsequently be annealed and reamplified to generate a chimeric gene sequence (see, for example, Ausubel et. al., (eds.) CURRENT PROTOCOLS IN MOLECULAR BIOLOGY, John Wiley & Sons, 1992). Moreover, many expression vectors are commercially available that already encode a fusion moiety (*e.g.*, a GST polypeptide). A nucleic acid encoding a polypeptide of the invention can be cloned into such an expression vector such that the fusion moiety is linked in-frame to the protein of the invention.

5.8 GENE THERAPY

Mutations in the polynucleotides of the invention gene may result in loss of normal function of the encoded protein. The invention thus provides gene therapy to restore normal activity of the polypeptides of the invention; or to treat disease states involving polypeptides of the invention. Delivery of a functional gene encoding polypeptides of the invention to appropriate cells is effected *ex vivo*, *in situ*, or *in vivo* by use of vectors, and more particularly viral vectors (*e.g.*, adenovirus, adeno-associated virus, or a retrovirus), or *ex vivo* by use of physical DNA transfer methods (*e.g.*, liposomes or chemical treatments). See, for example, Anderson, Nature, supplement to vol. 392, no. 6679, pp. 25-20 (1998). For additional reviews of gene therapy technology see Friedmann, Science, 244: 1275-1281 (1989); Verma, Scientific American: 68-84 (1990); and Miller, Nature, 357: 455-460 (1992). Introduction of any one of the nucleotides of the present invention or a gene encoding the polypeptides of the present invention can also be accomplished with extrachromosomal substrates (transient

expression) or artificial chromosomes (stable expression). Cells may also be cultured *ex vivo* in the presence of proteins of the present invention in order to proliferate or to produce a desired effect on or activity in such cells. Treated cells can then be introduced *in vivo* for therapeutic purposes. Alternatively, it is contemplated that in other human disease states, preventing the expression of or inhibiting the activity of polypeptides of the invention will be useful in treating the disease states. It is contemplated that antisense therapy or gene therapy could be applied to negatively regulate the expression of polypeptides of the invention.

Other methods inhibiting expression of a protein include the introduction of antisense molecules to the nucleic acids of the present invention, their complements, or their translated RNA sequences, by methods known in the art. Further, the polypeptides of the present invention can be inhibited by using targeted deletion methods, or the insertion of a negative regulatory element such as a silencer, which is tissue specific.

The present invention still further provides cells genetically engineered *in vivo* to express the polynucleotides of the invention, wherein such polynucleotides are in operative association with a regulatory sequence heterologous to the host cell which drives expression of the polynucleotides in the cell. These methods can be used to increase or decrease the expression of the polynucleotides of the present invention.

Knowledge of DNA sequences provided by the invention allows for modification of cells to permit, increase, or decrease, expression of endogenous polypeptide. Cells can be modified (*e.g.*, by homologous recombination) to provide increased polypeptide expression by replacing, in whole or in part, the naturally occurring promoter with all or part of a heterologous promoter so that the cells express the protein at higher levels. The heterologous promoter is inserted in such a manner that it is operatively linked to the desired protein encoding sequences. See, for example, PCT International Publication No. WO 94/12650, PCT International Publication No. WO 92/20808, and PCT International Publication No. WO 91/09955. It is also contemplated that, in addition to heterologous promoter DNA, amplifiable marker DNA (*e.g.*, ada, dhfr, and the multifunctional CAD gene which encodes carbamyl phosphate synthase, aspartate transcarbamylase, and dihydroorotase) and/or intron DNA may be inserted along with the heterologous promoter DNA. If linked to the desired

protein coding sequence, amplification of the marker DNA by standard selection methods results in co-amplification of the desired protein coding sequences in the cells.

In another embodiment of the present invention, cells and tissues may be engineered to express an endogenous gene comprising the polynucleotides of the invention under the control of inducible regulatory elements, in which case the regulatory sequences of the endogenous gene may be replaced by homologous recombination. As described herein, gene targeting can be used to replace a gene's existing regulatory region with a regulatory sequence isolated from a different gene or a novel regulatory sequence synthesized by genetic engineering methods. Such regulatory sequences may be comprised of promoters, enhancers, scaffold-attachment regions, negative regulatory elements, transcriptional initiation sites, regulatory protein binding sites or combinations of said sequences. Alternatively, sequences which affect the structure or stability of the RNA or protein produced may be replaced, removed, added, or otherwise modified by targeting. These sequences include polyadenylation signals, mRNA stability elements, splice sites, leader sequences for enhancing or modifying transport or secretion properties of the protein, or other sequences which alter or improve the function or stability of protein or RNA molecules.

The targeting event may be a simple insertion of the regulatory sequence, placing the gene under the control of the new regulatory sequence, *e.g.*, inserting a new promoter or enhancer or both upstream of a gene. Alternatively, the targeting event may be a simple deletion of a regulatory element, such as the deletion of a tissue-specific negative regulatory element. Alternatively, the targeting event may replace an existing element; for example, a tissue-specific enhancer can be replaced by an enhancer that has broader or different cell-type specificity than the naturally occurring elements. Here, the naturally occurring sequences are deleted and new sequences are added. In all cases, the identification of the targeting event may be facilitated by the use of one or more selectable marker genes that are contiguous with the targeting DNA, allowing for the selection of cells in which the exogenous DNA has integrated into the cell genome. The identification of the targeting event may also be facilitated by the use of one or more marker genes exhibiting the property of negative selection, such that the negatively selectable marker is linked to the exogenous DNA, but configured such that the negatively selectable marker flanks the targeting

sequence, and such that a correct homologous recombination event with sequences in the host cell genome does not result in the stable integration of the negatively selectable marker. Markers useful for this purpose include the Herpes Simplex Virus thymidine kinase (TK) gene or the bacterial xanthine-guanine phosphoribosyl-transferase (gpt) gene.

The gene targeting or gene activation techniques which can be used in accordance with this aspect of the invention are more particularly described in U.S. Patent No. 5,272,071 to Chappel; U.S. Patent No. 5,578,461 to Sherwin et. al.,; International Application No. PCT/US92/09627 (WO93/09222) by Selden et. al.,; and International Application No. PCT/US90/06436 (WO91/06667) by Skoultchi *et al.*, each of which is incorporated by reference herein in its entirety.

5.9 TRANSGENIC ANIMALS

In preferred methods to determine biological functions of the polypeptides of the invention *in vivo*, one or more genes provided by the invention are either over expressed or inactivated in the germ line of animals using homologous recombination (Capecchi, Science 244:1288-1292 (1989)). Animals in which the gene is over expressed, under the regulatory control of exogenous or endogenous promoter elements, are known as transgenic animals. Animals in which an endogenous gene has been inactivated by homologous recombination are referred to as "knockout" animals. Knockout animals, preferably non-human mammals, can be prepared as described in U.S. Patent No. 5,557,032, incorporated herein by reference. Transgenic animals are useful to determine the roles polypeptides of the invention play in biological processes, and preferably in disease states. Transgenic animals are useful as model systems to identify compounds that modulate lipid metabolism. Transgenic animals, preferably non-human mammals, are produced using methods as described in U.S. Patent No 5,489,743 and PCT Publication No. WO94/28122, incorporated herein by reference.

Transgenic animals can be prepared wherein all or part of a promoter of the polynucleotides of the invention is either activated or inactivated to alter the level of expression of the polypeptides of the invention. Inactivation can be carried out using homologous recombination methods described above. Activation can be achieved by supplementing or even replacing the homologous promoter to provide for increased

protein expression. The homologous promoter can be supplemented by insertion of one or more heterologous enhancer elements known to confer promoter activation in a particular tissue.

The polynucleotides of the present invention also make possible the development, through, *e.g.*, homologous recombination or knock out strategies, of animals that fail to express polypeptides of the invention or that express a variant polypeptide. Such animals are useful as models for studying the *in vivo* activities of polypeptide as well as for studying modulators of the polypeptides of the invention.

In preferred methods to determine biological functions of the polypeptides of the invention *in vivo*, one or more genes provided by the invention are either over expressed or inactivated in the germ line of animals using homologous recombination (Capecchi, Science 244:1288-1292 (1989)). Animals in which the gene is over expressed, under the regulatory control of exogenous or endogenous promoter elements, are known as transgenic animals. Animals in which an endogenous gene has been inactivated by homologous recombination are referred to as "knockout" animals. Knockout animals, preferably non-human mammals, can be prepared as described in U.S. Patent No. 5,557,032, incorporated herein by reference. Transgenic animals are useful to determine the roles polypeptides of the invention play in biological processes, and preferably in disease states. Transgenic animals are useful as model systems to identify compounds that modulate lipid metabolism. Transgenic animals, preferably non-human mammals, are produced using methods as described in U.S. Patent No 5,489,743 and PCT Publication No. WO94/28122, incorporated herein by reference.

Transgenic animals can be prepared wherein all or part of the polynucleotides of the invention promoter is either activated or inactivated to alter the level of expression of the polypeptides of the invention. Inactivation can be carried out using homologous recombination methods described above. Activation can be achieved by supplementing or even replacing the homologous promoter to provide for increased protein expression. The homologous promoter can be supplemented by insertion of one or more heterologous enhancer elements known to confer promoter activation in a particular tissue.

5.10 USES AND BIOLOGICAL ACTIVITY

The polynucleotides and proteins of the present invention are expected to exhibit one or more of the uses or biological activities (including those associated with assays cited herein) identified herein. Uses or activities described for proteins of the present invention may be provided by administration or use of such proteins or of polynucleotides encoding such proteins (such as, for example, in gene therapies or vectors suitable for introduction of DNA). The mechanism underlying the particular condition or pathology will dictate whether the polypeptides of the invention, the polynucleotides of the invention or modulators (activators or inhibitors) thereof would be beneficial to the subject in need of treatment. Thus, "therapeutic compositions of the invention" include compositions comprising isolated polynucleotides (including recombinant DNA molecules, cloned genes and degenerate variants thereof) or polypeptides of the invention (including full length protein, mature protein and truncations or domains thereof), or compounds and other substances that modulate the overall activity of the target gene products, either at the level of target gene/protein expression or target protein activity. Such modulators include polypeptides, analogs, (variants), including fragments and fusion proteins, antibodies and other binding proteins; chemical compounds that directly or indirectly activate or inhibit the polypeptides of the invention (identified, *e.g.,* via drug screening assays as described herein); antisense polynucleotides and polynucleotides suitable for triple helix formation; and in particular antibodies or other binding partners that specifically recognize one or more epitopes of the polypeptides of the invention.

The polypeptides of the present invention may likewise be involved in cellular activation or in one of the other physiological pathways described herein.

5.10.1 RESEARCH USES AND UTILITIES

The polynucleotides provided by the present invention can be used by the research community for various purposes. The polynucleotides can be used to express recombinant protein for analysis, characterization or therapeutic use; as markers for tissues in which the corresponding protein is preferentially expressed (either constitutively or at a particular stage of tissue differentiation or development or in disease states); as molecular weight markers on gels; as chromosome markers or tags (when

labeled) to identify chromosomes or to map related gene positions; to compare with endogenous DNA sequences in patients to identify potential genetic disorders; as probes to hybridize and thus discover novel, related DNA sequences; as a source of information to derive PCR primers for genetic fingerprinting; as a probe to "subtract-out" known sequences in the process of discovering other novel polynucleotides; for selecting and making oligomers for attachment to a "gene chip" or other support, including for examination of expression patterns; to raise anti-protein antibodies using DNA immunization techniques; and as an antigen to raise anti-DNA antibodies or elicit another immune response. Where the polynucleotide encodes a protein which binds or potentially binds to another protein (such as, for example, in a receptor-ligand interaction), the polynucleotide can also be used in interaction trap assays (such as, for example, that described in Gyuris *et al.*, Cell 75:791-803 (1993)) to identify polynucleotides encoding the other protein with which binding occurs or to identify inhibitors of the binding interaction.

The polypeptides provided by the present invention can similarly be used in assays to determine biological activity, including in a panel of multiple proteins for high-throughput screening; to raise antibodies or to elicit another immune response; as a reagent (including the labeled reagent) in assays designed to quantitatively determine levels of the protein (or its receptor) in biological fluids; as markers for tissues in which the corresponding polypeptide is preferentially expressed (either constitutively or at a particular stage of tissue differentiation or development or in a disease state); and, of course, to isolate correlative receptors or ligands. Proteins involved in these binding interactions can also be used to screen for peptide or small molecule inhibitors or agonists of the binding interaction.

Any or all of these research utilities are capable of being developed into reagent grade or kit format for commercialization as research products.

Methods for performing the uses listed above are well known to those skilled in the art. References disclosing such methods include without limitation "Molecular Cloning: A Laboratory Manual", 2d ed., Cold Spring Harbor Laboratory Press, Sambrook, J., E. F. Fritsch and T. Maniatis eds., 1989, and "Methods in Enzymology:

Guide to Molecular Cloning Techniques", Academic Press, Berger, S. L. and A. R. Kimmel eds., 1987.

5.10.2 NUTRITIONAL USES

Polynucleotides and polypeptides of the present invention can also be used as nutritional sources or supplements. Such uses include without limitation use as a protein or amino acid supplement, use as a carbon source, use as a nitrogen source and use as a source of carbohydrate. In such cases the polypeptide or polynucleotide of the invention can be added to the feed of a particular organism or can be administered as a separate solid or liquid preparation, such as in the form of powder, pills, solutions, suspensions or capsules. In the case of microorganisms, the polypeptide or polynucleotide of the invention can be added to the medium in or on which the microorganism is cultured.

5.10.3 CYTOKINE AND CELL PROLIFERATION/DIFFERENTIATION ACTIVITY

A polypeptide of the present invention may exhibit activity relating to cytokine, cell proliferation (either inducing or inhibiting) or cell differentiation (either inducing or inhibiting) activity or may induce production of other cytokines in certain cell populations. A polynucleotide of the invention can encode a polypeptide exhibiting such attributes. Many protein factors discovered to date, including all known cytokines, have exhibited activity in one or more factor-dependent cell proliferation assays, and hence the assays serve as a convenient confirmation of cytokine activity. The activity of therapeutic compositions of the present invention is evidenced by any one of a number of routine factor dependent cell proliferation assays for cell lines including, without limitation, 32D, DA2, DA1G, T10, B9, B9/11, BaF3, MC9/G, M+(preB M+), 2E8, RB5, DA1, 123, T1165, HT2, CTLL2, TF-1, Mo7e, CMK, HUVEC, and CaCo. Therapeutic compositions of the invention can be used in the following:

Assays for T-cell or thymocyte proliferation include without limitation those described in: Current Protocols in Immunology, Ed by J. E. Coligan, A. M. Kruisbeek, D. H. Margulies, E. M. Shevach, W. Strober, Pub. Greene Publishing Associates and Wiley-Interscience (Chapter 3, *In vitro* assays for Mouse Lymphocyte Function 3.1-3.19;

Chapter 7, Immunologic studies in Humans); Takai *et al.*, J. Immunol. 137:3494-3500, 1986; Bertagnolli *et al.*, J. Immunol. 145:1706-1712, 1990; Bertagnolli *et al.*, Cellular Immunology 133:327-341, 1991; Bertagnolli, *et al.*, I. Immunol. 149:3778-3783, 1992; Bowman *et al.*, I. Immunol. 152:1756-1761, 1994.

Assays for cytokine production and/or proliferation of spleen cells, lymph node cells or thymocytes include, without limitation, those described in: Polyclonal T cell stimulation, Kruisbeek, A. M. and Shevach, E. M. In Current Protocols in Immunology. J. E. e.a. Coligan eds. Vol 1 pp. 3.12.1-3.12.14, John Wiley and Sons, Toronto. 1994; and Measurement of mouse and human interleukin-γ, Schreiber, R. D. In Current Protocols in Immunology. J. E. e.a. Coligan eds. Vol 1 pp. 6.8.1-6.8.8, John Wiley and Sons, Toronto. 1994.

Assays for proliferation and differentiation of hematopoietic and lymphopoietic cells include, without limitation, those described in: Measurement of Human and Murine Interleukin 2 and Interleukin 4, Bottomly, K., Davis, L. S. and Lipsky, P. E. In Current Protocols in Immunology. J. E. e.a. Coligan eds. Vol 1 pp. 6.3.1-6.3.12, John Wiley and Sons, Toronto. 1991; deVries *et al.*, J. Exp. Med. 173:1205-1211, 1991; Moreau *et al.*, Nature 336:690-692, 1988; Greenberger *et al.*, Proc. Natl. Acad. Sci. U.S.A. 80:2931-2938, 1983; Measurement of mouse and human interleukin 6--Nordan, R. In Current Protocols in Immunology. J. E. Coligan eds. Vol 1 pp. 6.6.1-6.6.5, John Wiley and Sons, Toronto. 1991; Smith *et al.*, Proc. Natl. Aced. Sci. U.S.A. 83:1857-1861, 1986; Measurement of human Interleukin 11--Bennett, F., Giannotti, J., Clark, S. C. and Turner, K. J. In Current Protocols in Immunology. J. E. Coligan eds. Vol 1 pp. 6.15.1 John Wiley and Sons, Toronto. 1991; Measurement of mouse and human Interleukin 9--Ciarletta, A., Giannotti, J., Clark, S. C. and Turner, K. J. In Current Protocols in Immunology. J. E. Coligan eds. Vol 1 pp. 6.13.1, John Wiley and Sons, Toronto. 1991.

Assays for T-cell clone responses to antigens (which will identify, among others, proteins that affect APC-T cell interactions as well as direct T-cell effects by measuring proliferation and cytokine production) include, without limitation, those described in: Current Protocols in Immunology, Ed by J. E. Coligan, A. M. Kruisbeek, D. H. Margulies, E. M. Shevach, W Strober, Pub. Greene Publishing Associates and Wiley-Interscience (Chapter 3, *In vitro* assays for Mouse Lymphocyte Function; Chapter

6, Cytokines and their cellular receptors; Chapter 7, Immunologic studies in Humans); Weinberger *et al.*, Proc. Natl. Acad. Sci. USA 77:6091-6095, 1980; Weinberger *et al.*, Eur. J. Immun. 11:405-411, 1981; Takai *et al.*, J. Immunol. 137:3494-3500, 1986; Takai *et al.*, J. Immunol. 140:508-512, 1988.

5.10.4 STEM CELL GROWTH FACTOR ACTIVITY

A polypeptide of the present invention may exhibit stem cell growth factor activity and be involved in the proliferation, differentiation and survival of pluripotent and totipotent stem cells including primordial germ cells, embryonic stem cells, hematopoietic stem cells and/or germ line stem cells. Administration of the polypeptide of the invention to stem cells *in vivo* or *ex vivo* is expected to maintain and expand cell populations in a totipotential or pluripotential state which would be useful for re-engineering damaged or diseased tissues, transplantation, manufacture of bio-pharmaceuticals and the development of bio-sensors. The ability to produce large quantities of human cells has important working applications for the production of human proteins which currently must be obtained from non-human sources or donors, implantation of cells to treat diseases such as Parkinson's, Alzheimer's and other neurodegenerative diseases; tissues for grafting such as bone marrow, skin, cartilage, tendons, bone, muscle (including cardiac muscle), blood vessels, cornea, neural cells, gastrointestinal cells and others; and organs for transplantation such as kidney, liver, pancreas (including islet cells), heart and lung.

It is contemplated that multiple different exogenous growth factors and/or cytokines may be administered in combination with the polypeptide of the invention to achieve the desired effect, including any of the growth factors listed herein, other stem cell maintenance factors, and specifically including stem cell factor (SCF), leukemia inhibitory factor (LIF), Flt-3 ligand (Flt-3L), any of the interleukins, recombinant soluble IL-6 receptor fused to IL-6, macrophage inflammatory protein 1-alpha (MIP-1-alpha), G-CSF, GM-CSF, thrombopoietin (TPO), platelet factor 4 (PF-4), platelet-derived growth factor (PDGF), neural growth factors and basic fibroblast growth factor (bFGF).

Since totipotent stem cells can give rise to virtually any mature cell type, expansion of these cells in culture will facilitate the production of large quantities of

mature cells. Techniques for culturing stem cells are known in the art and administration of polypeptides of the invention, optionally with other growth factors and/or cytokines, is expected to enhance the survival and proliferation of the stem cell populations. This can be accomplished by direct administration of the polypeptide of the invention to the culture medium. Alternatively, stroma cells transfected with a polynucleotide that encodes for the polypeptide of the invention can be used as a feeder layer for the stem cell populations in culture or *in vivo*. Stromal support cells for feeder layers may include embryonic bone marrow fibroblasts, bone marrow stromal cells, fetal liver cells, or cultured embryonic fibroblasts (see U.S. Patent No. 5,690,926).

Stem cells themselves can be transfected with a polynucleotide of the invention to induce autocrine expression of the polypeptide of the invention. This will allow for generation of undifferentiated totipotential/pluripotential stem cell lines that are useful as is or that can then be differentiated into the desired mature cell types. These stable cell lines can also serve as a source of undifferentiated totipotential/pluripotential mRNA to create cDNA libraries and templates for polymerase chain reaction experiments. These studies would allow for the isolation and identification of differentially expressed genes in stem cell populations that regulate stem cell proliferation and/or maintenance.

Expansion and maintenance of totipotent stem cell populations will be useful in the treatment of many pathological conditions. For example, polypeptides of the present invention may be used to manipulate stem cells in culture to give rise to neuroepithelial cells that can be used to augment or replace cells damaged by illness, autoimmune disease, accidental damage or genetic disorders. The polypeptide of the invention may be useful for inducing the proliferation of neural cells and for the regeneration of nerve and brain tissue, i.e. for the treatment of central and peripheral nervous system diseases and neuropathies, as well as mechanical and traumatic disorders which involve degeneration, death or trauma to neural cells or nerve tissue. In addition, the expanded stem cell populations can also be genetically altered for gene therapy purposes and to decrease host rejection of replacement tissues after grafting or implantation.

Expression of the polypeptide of the invention and its effect on stem cells can also be manipulated to achieve controlled differentiation of the stem cells into more differentiated cell types. A broadly applicable method of obtaining pure populations of a

specific differentiated cell type from undifferentiated stem cell populations involves the use of a cell-type specific promoter driving a selectable marker. The selectable marker allows only cells of the desired type to survive. For example, stem cells can be induced to differentiate into cardiomyocytes (Wobus *et al.*, Differentiation, 48: 173-182, (1991); Klug *et al.*, J. Clin. Invest., 98(1): 216-224, (1998)) or skeletal muscle cells (Browder, L. W. In: *Principles of Tissue Engineering eds.* Lanza *et al.*, Academic Press (1997)). Alternatively, directed differentiation of stem cells can be accomplished by culturing the stem cells in the presence of a differentiation factor such as retinoic acid and an antagonist of the polypeptide of the invention which would inhibit the effects of endogenous stem cell factor activity and allow differentiation to proceed.

In vitro cultures of stem cells can be used to determine if the polypeptide of the invention exhibits stem cell growth factor activity. Stem cells are isolated from any one of various cell sources (including hematopoietic stem cells and embryonic stem cells) and cultured on a feeder layer, as described by Thompson et. al., Proc. Natl. Acad. Sci, U.S.A., 92: 7844-7848 (1995), in the presence of the polypeptide of the invention alone or in combination with other growth factors or cytokines. The ability of the polypeptide of the invention to induce stem cells proliferation is determined by colony formation on semi-solid support e.g. as described by Bernstein Blood, 77: 2316-2321 (1991).

5.10.5 HEMATOPOIESIS REGULATING ACTIVITY

A polypeptide of the present invention may be involved in regulation of hematopoiesis and, consequently, in the treatment of myeloid or lymphoid cell disorders. Even marginal biological activity in support of colony forming cells or of factor-dependent cell lines indicates involvement in regulating hematopoiesis, e.g. in supporting the growth and proliferation of erythroid progenitor cells alone or in combination with other cytokines, thereby indicating utility, for example, in treating various anemias or for use in conjunction with irradiation/chemotherapy to stimulate the production of erythroid precursors and/or erythroid cells; in supporting the growth and proliferation of myeloid cells such as granulocytes and monocytes/macrophages (*i.e.*, traditional CSF activity) useful, for example, in conjunction with chemotherapy to prevent or treat consequent myelo-suppression; in supporting the growth and proliferation

of megakaryocytes and consequently of platelets thereby allowing prevention or treatment of various platelet disorders such as thrombocytopenia, and generally for use in place of or complimentary to platelet transfusions; and/or in supporting the growth and proliferation of hematopoietic stem cells which are capable of maturing to any and all of the above-mentioned hematopoietic cells and therefore find therapeutic utility in various stem cell disorders (such as those usually treated with transplantation, including, without limitation, aplastic anemia and paroxysmal nocturnal hemoglobinuria), as well as in repopulating the stem cell compartment post irradiation/chemotherapy, either *in-vivo* or *ex-vivo* (*i.e.,* in conjunction with bone marrow transplantation or with peripheral progenitor cell transplantation (homologous or heterologous)) as normal cells or genetically manipulated for gene therapy.

Therapeutic compositions of the invention can be used in the following:

Suitable assays for proliferation and differentiation of various hematopoietic lines are cited above.

Assays for embryonic stem cell differentiation (which will identify, among others, proteins that influence embryonic differentiation hematopoiesis) include, without limitation, those described in: Johansson et. al., Cellular Biology 15:141-151, 1995; Keller et al.. Molecular and Cellular Biology 13:473-486, 1993; McClanahan et al.. Blood 81:2903-2915, 1993.

Assays for stem cell survival and differentiation (which will identify, among others, proteins that regulate lympho-hematopoiesis) include, without limitation, those described in: Methylcellulose colony forming assays, Freshney, M. G. In Culture of Hematopoietic Cells. R. I. Freshney, et. al., eds. Vol pp. 265-268, Wiley-Liss, Inc., New York, N.Y. 1994; Hirayama et al.. Proc. Natl. Acad. Sci. USA 89:5907-5911, 1992; Primitive hematopoietic colony forming cells with high proliferative potential, McNiece, I. K. and Briddell, R. A. In Culture of Hematopoietic Cells. R. I. Freshney, et. al., eds. Vol pp. 23-39, Wiley-Liss, Inc., New York, N.Y. 1994; Neben et al.. Experimental Hematology 22:353-359, 1994; Cobblestone area forming cell assay, Ploemacher, R. E. In Culture of Hematopoietic Cells. R. I. Freshney, et. al., eds. Vol pp. 1-21, Wiley-Liss, Inc., New York, N.Y. 1994; Long term bone marrow cultures in the presence of stromal cells, Spooncer, E., Dexter, M. and Allen, T. In Culture of Hematopoietic Cells. R. I.

Freshney, et. al., eds. Vol pp. 163-179, Wiley-Liss, Inc., New York, N.Y. 1994; Long term culture initiating cell assay, Sutherland, H. J. In Culture of Hematopoietic Cells. R. I. Freshney, et. al., eds. Vol pp. 139-162, Wiley-Liss, Inc., New York, N.Y. 1994.

5.10.6 TISSUE GROWTH ACTIVITY

A polypeptide of the present invention also may be involved in bone, cartilage, tendon, ligament and/or nerve tissue growth or regeneration, as well as in wound healing and tissue repair and replacement, and in healing of burns, incisions and ulcers.

A polypeptide of the present invention which induces cartilage and/or bone growth in circumstances where bone is not normally formed, has application in the healing of bone fractures and cartilage damage or defects in humans and other animals. Compositions of a polypeptide, antibody, binding partner, or other modulator of the invention may have prophylactic use in closed as well as open fracture reduction and also in the improved fixation of artificial joints. *De novo* bone formation induced by an osteogenic agent contributes to the repair of congenital, trauma induced, or oncologic resection induced craniofacial defects, and also is useful in cosmetic plastic surgery.

A polypeptide of this invention may also be involved in attracting bone-forming cells, stimulating growth of bone-forming cells, or inducing differentiation of progenitors of bone-forming cells. Treatment of osteoporosis, osteoarthritis, bone degenerative disorders, or periodontal disease, such as through stimulation of bone and/or cartilage repair or by blocking inflammation or processes of tissue destruction (collagenase activity, osteoclast activity, etc.) mediated by inflammatory processes may also be possible using the composition of the invention.

Another category of tissue regeneration activity that may involve the polypeptide of the present invention is tendon/ligament formation. Induction of tendon/ligament-like tissue or other tissue formation in circumstances where such tissue is not normally formed, has application in the healing of tendon or ligament tears, deformities and other tendon or ligament defects in humans and other animals. Such a preparation employing a tendon/ligament-like tissue inducing protein may have prophylactic use in preventing damage to tendon or ligament tissue, as well as use in the improved fixation of tendon or ligament to bone or other tissues, and in repairing defects to tendon or ligament tissue.

De novo tendon/ligament-like tissue formation induced by a composition of the present invention contributes to the repair of congenital, trauma induced, or other tendon or ligament defects of other origin, and is also useful in cosmetic plastic surgery for attachment or repair of tendons or ligaments. The compositions of the present invention may provide environment to attract tendon- or ligament-forming cells, stimulate growth of tendon- or ligament-forming cells, induce differentiation of progenitors of tendon- or ligament-forming cells, or induce growth of tendon/ligament cells or progenitors *ex vivo* for return *in vivo* to effect tissue repair. The compositions of the invention may also be useful in the treatment of tendinitis, carpal tunnel syndrome and other tendon or ligament defects. The compositions may also include an appropriate matrix and/or sequestering agent as a carrier as is well known in the art.

The compositions of the present invention may also be useful for proliferation of neural cells and for regeneration of nerve and brain tissue, i.e. for the treatment of central and peripheral nervous system diseases and neuropathies, as well as mechanical and traumatic disorders, which involve degeneration, death or trauma to neural cells or nerve tissue. More specifically, a composition may be used in the treatment of diseases of the peripheral nervous system, such as peripheral nerve injuries, peripheral neuropathy and localized neuropathies, and central nervous system diseases, such as Alzheimer's, Parkinson's disease, Huntington's disease, amyotrophic lateral sclerosis, and Shy-Drager syndrome. Further conditions which may be treated in accordance with the present invention include mechanical and traumatic disorders, such as spinal cord disorders, head trauma and cerebrovascular diseases such as stroke. Peripheral neuropathies resulting from chemotherapy or other medical therapies may also be treatable using a composition of the invention.

Compositions of the invention may also be useful to promote better or faster closure of non-healing wounds, including without limitation pressure ulcers, ulcers associated with vascular insufficiency, surgical and traumatic wounds, and the like.

Compositions of the present invention may also be involved in the generation or regeneration of other tissues, such as organs (including, for example, pancreas, liver, intestine, kidney, skin, endothelium), muscle (smooth, skeletal or cardiac) and vascular (including vascular endothelium) tissue, or for promoting the growth of cells comprising

such tissues. Part of the desired effects may be by inhibition or modulation of fibrotic scarring may allow normal tissue to regenerate. A polypeptide of the present invention may also exhibit angiogenic activity.

A composition of the present invention may also be useful for gut protection or regeneration and treatment of lung or liver fibrosis, reperfusion injury in various tissues, and conditions resulting from systemic cytokine damage.

A composition of the present invention may also be useful for promoting or inhibiting differentiation of tissues described above from precursor tissues or cells; or for inhibiting the growth of tissues described above.

Therapeutic compositions of the invention can be used in the following:

Assays for tissue generation activity include, without limitation, those described in: International Patent Publication No. WO95/16035 (bone, cartilage, tendon); International Patent Publication No. WO95/05846 (nerve, neuronal); International Patent Publication No. WO91/07491 (skin, endothelium).

Assays for wound healing activity include, without limitation, those described in: Winter, Epidermal Wound Healing, pps. 71-112 (Maibach, H. I. and Rovee, D. T., eds.), Year Book Medical Publishers, Inc., Chicago, as modified by Eaglstein and Mertz, J. Invest. Dermatol 71:382-84 (1978).

5.10.7 IMMUNE STIMULATING OR SUPPRESSING ACTIVITY

A polypeptide of the present invention may also exhibit immune stimulating or immune suppressing activity, including without limitation the activities for which assays are described herein. A polynucleotide of the invention can encode a polypeptide exhibiting such activities. A protein may be useful in the treatment of various immune deficiencies and disorders (including severe combined immunodeficiency (SCID)), *e.g.*, in regulating (up or down) growth and proliferation of T and/or B lymphocytes, as well as effecting the cytolytic activity of NK cells and other cell populations. These immune deficiencies may be genetic or be caused by viral (*e.g.*, HIV) as well as bacterial or fungal infections, or may result from autoimmune disorders. More specifically, infectious diseases causes by viral, bacterial, fungal or other infection may be treatable using a protein of the present invention, including infections by HIV, hepatitis viruses, herpes

viruses, mycobacteria, Leishmania spp., malaria spp. and various fungal infections such as candidiasis. Of course, in this regard, proteins of the present invention may also be useful where a boost to the immune system generally may be desirable, *i.e.*, in the treatment of cancer.

Autoimmune disorders which may be treated using a protein of the present invention include, for example, connective tissue disease, multiple sclerosis, systemic lupus erythematosus, rheumatoid arthritis, autoimmune pulmonary inflammation, Guillain-Barre syndrome, autoimmune thyroiditis, insulin dependent diabetes mellitis, myasthenia gravis, graft-versus-host disease and autoimmune inflammatory eye disease. Such a protein (or antagonists thereof, including antibodies) of the present invention may also to be useful in the treatment of allergic reactions and conditions (*e.g.*, anaphylaxis, serum sickness, drug reactions, food allergies, insect venom allergies, mastocytosis, allergic rhinitis, hypersensitivity pneumonitis, urticaria, angioedema, eczema, atopic dermatitis, allergic contact dermatitis, erythema multiforme, Stevens-Johnson syndrome, allergic conjunctivitis, atopic keratoconjunctivitis, venereal keratoconjunctivitis, giant papillary conjunctivitis and contact allergies), such as asthma (particularly allergic asthma) or other respiratory problems. Other conditions, in which immune suppression is desired (including, for example, organ transplantation), may also be treatable using a protein (or antagonists thereof) of the present invention. The therapeutic effects of the polypeptides or antagonists thereof on allergic reactions can be evaluated by *in vivo* animals models such as the cumulative contact enhancement test (Lastbom et al.. Toxicology 125: 59-66, 1998), skin prick test (Hoffmann et al.. Allergy 54: 446-54, 1999), guinea pig skin sensitization test (Vohr et al.. Arch. Toxocol. 73: 501-9), and murine local lymph node assay (Kimber et al.. J. Toxicol. Environ. Health 53: 563-79).

Using the proteins of the invention it may also be possible to modulate immune responses, in a number of ways. Down regulation may be in the form of inhibiting or blocking an immune response already in progress or may involve preventing the induction of an immune response. The functions of activated T cells may be inhibited by suppressing T cell responses or by inducing specific tolerance in T cells, or both. Immunosuppression of T cell responses is generally an active, non-antigen-specific, process which requires continuous exposure of the T cells to the suppressive agent.

Tolerance, which involves inducing non-responsiveness or anergy in T cells, is distinguishable from immunosuppression in that it is generally antigen-specific and persists after exposure to the tolerizing agent has ceased. Operationally, tolerance can be demonstrated by the lack of a T cell response upon reexposure to specific antigen in the absence of the tolerizing agent.

Down regulating or preventing one or more antigen functions (including without limitation B lymphocyte antigen functions (such as, for example, B7)), *e.g.,* preventing high level lymphokine synthesis by activated T cells, will be useful *in situ*ations of tissue, skin and organ transplantation and in graft-versus-host disease (GVHD). For example, blockage of T cell function should result in reduced tissue destruction in tissue transplantation. Typically, in tissue transplants, rejection of the transplant is initiated through its recognition as foreign by T cells, followed by an immune reaction that destroys the transplant. The administration of a therapeutic composition of the invention may prevent cytokine synthesis by immune cells, such as T cells, and thus acts as an immunosuppressant. Moreover, a lack of costimulation may also be sufficient to anergize the T cells, thereby inducing tolerance in a subject. Induction of long-term tolerance by B lymphocyte antigen-blocking reagents may avoid the necessity of repeated administration of these blocking reagents. To achieve sufficient immunosuppression or tolerance in a subject, it may also be necessary to block the function of a combination of B lymphocyte antigens.

The efficacy of particular therapeutic compositions in preventing organ transplant rejection or GVHD can be assessed using animal models that are predictive of efficacy in humans. Examples of appropriate systems which can be used include allogeneic cardiac grafts in rats and xenogeneic pancreatic islet cell grafts in mice, both of which have been used to examine the immunosuppressive effects of CTLA4Ig fusion proteins *in vivo* as described in Lenschow et al.. Science 257:789-792 (1992) and Turka et al.. Proc. Natl. Acad. Sci USA, 89:11102-11105 (1992). In addition, murine models of GVHD (see Paul ed., Fundamental Immunology, Raven Press, New York, 1989, pp. 846-847) can be used to determine the effect of therapeutic compositions of the invention on the development of that disease.

Blocking antigen function may also be therapeutically useful for treating autoimmune diseases. Many autoimmune disorders are the result of inappropriate activation of T cells that are reactive against self tissue and which promote the production of cytokines and autoantibodies involved in the pathology of the diseases. Preventing the activation of autoreactive T cells may reduce or eliminate disease symptoms. Administration of reagents which block stimulation of T cells can be used to inhibit T cell activation and prevent production of autoantibodies or T cell-derived cytokines which may be involved in the disease process. Additionally, blocking reagents may induce antigen-specific tolerance of autoreactive T cells which could lead to long-term relief from the disease. The efficacy of blocking reagents in preventing or alleviating autoimmune disorders can be determined using a number of well-characterized animal models of human autoimmune diseases. Examples include murine experimental autoimmune encephalitis, systemic lupus erythmatosis in MRL/lpr/lpr mice or NZB hybrid mice, murine autoimmune collagen arthritis, diabetes mellitus in NOD mice and BB rats, and murine experimental myasthenia gravis (see Paul ed., Fundamental Immunology, Raven Press, New York, 1989, pp. 840-856).

Upregulation of an antigen function (*e.g.*, a B lymphocyte antigen function), as a means of up regulating immune responses, may also be useful in therapy. Upregulation of immune responses may be in the form of enhancing an existing immune response or eliciting an initial immune response. For example, enhancing an immune response may be useful in cases of viral infection, including systemic viral diseases such as influenza, the common cold, and encephalitis.

Alternatively, anti-viral immune responses may be enhanced in an infected patient by removing T cells from the patient, costimulating the T cells *in vitro* with viral antigen-pulsed APCs either expressing a peptide of the present invention or together with a stimulatory form of a soluble peptide of the present invention and reintroducing the *in vitro* activated T cells into the patient. Another method of enhancing anti-viral immune responses would be to isolate infected cells from a patient, transfect them with a nucleic acid encoding a protein of the present invention as described herein such that the cells express all or a portion of the protein on their surface, and reintroduce the transfected

cells into the patient. The infected cells would now be capable of delivering a costimulatory signal to, and thereby activate, T cells *in vivo*.

A polypeptide of the present invention may provide the necessary stimulation signal to T cells to induce a T cell mediated immune response against the transfected tumor cells. In addition, tumor cells which lack MHC class I or MHC class II molecules, or which fail to reexpress sufficient mounts of MHC class I or MHC class II molecules, can be transfected with nucleic acid encoding all or a portion of (*e.g.*, a cytoplasmic-domain truncated portion) of an MHC class I alpha chain protein and β_2 microglobulin protein or an MHC class II alpha chain protein and an MHC class II beta chain protein to thereby express MHC class I or MHC class II proteins on the cell surface. Expression of the appropriate class I or class II MHC in conjunction with a peptide having the activity of a B lymphocyte antigen (*e.g.*, B7-1, B7-2, B7-3) induces a T cell mediated immune response against the transfected tumor cell. Optionally, a gene encoding an antisense construct which blocks expression of an MHC class II associated protein, such as the invariant chain, can also be cotransfected with a DNA encoding a peptide having the activity of a B lymphocyte antigen to promote presentation of tumor associated antigens and induce tumor specific immunity. Thus, the induction of a T cell mediated immune response in a human subject may be sufficient to overcome tumor-specific tolerance in the subject.

The activity of a protein of the invention may, among other means, be measured by the following methods:

Suitable assays for thymocyte or splenocyte cytotoxicity include, without limitation, those described in: Current Protocols in Immunology, Ed by J. E. Coligan, A. M. Kruisbeek, D. H. Margulies, E. M. Shevach, W. Strober, Pub. Greene Publishing Associates and Wiley-Interscience (Chapter 3, *In Vitro* assays for Mouse Lymphocyte Function 3.1-3.19; Chapter 7, Immunologic studies in Humans); Herrmann et al.. Proc. Natl. Acad. Sci. USA 78:2488-2492, 1981; Herrmann et al.. J. Immunol. 128:1968-1974, 1982; Handa et al.. J. Immunol. 135:1564-1572, 1985; Takai et al.. I. Immunol. 137:3494-3500, 1986; Takai et al.. J. Immunol. 140:508-512, 1988; Bowman et al.. J. Virology 61:1992-1998; Bertagnolli et al.. Cellular Immunology 133:327-341, 1991; Brown et al.. J. Immunol. 153:3079-3092, 1994.

Assays for T-cell-dependent immunoglobulin responses and isotype switching (which will identify, among others, proteins that modulate T-cell dependent antibody responses and that affect Th1/Th2 profiles) include, without limitation, those described in: Maliszewski, J. Immunol. 144:3028-3033, 1990; and Assays for B cell function: *In vitro* antibody production, Mond, J. J. and Brunswick, M. In Current Protocols in Immunology. J. E. e.a. Coligan eds. Vol 1 pp. 3.8.1-3.8.16, John Wiley and Sons, Toronto. 1994.

Mixed lymphocyte reaction (MLR) assays (which will identify, among others, proteins that generate predominantly Th1 and CTL responses) include, without limitation, those described in: Current Protocols in Immunology, Ed by J. E. Coligan, A. M. Kruisbeek, D. H. Margulies, E. M. Shevach, W. Strober, Pub. Greene Publishing Associates and Wiley-Interscience (Chapter 3, *In vitro* assays for Mouse Lymphocyte Function 3.1-3.19; Chapter 7, Immunologic studies in Humans); Takai et al.. J. Immunol. 137:3494-3500, 1986; Takai et al.. J. Immunol. 140:508-512, 1988; Bertagnolli et al.. J. Immunol. 149:3778-3783, 1992.

Dendritic cell-dependent assays (which will identify, among others, proteins expressed by dendritic cells that activate naive T-cells) include, without limitation, those described in: Guery et al.. J. Immunol. 134:536-544, 1995; Inaba et al.. Journal of Experimental Medicine 173:549-559, 1991; Macatonia et al.. Journal of Immunology 154:5071-5079, 1995; Porgador et al.. Journal of Experimental Medicine 182:255-260, 1995; Nair et al.. Journal of Virology 67:4062-4069, 1993; Huang et al.. Science 264:961-965, 1994; Macatonia et al.. Journal of Experimental Medicine 169:1255-1264, 1989; Bhardwaj et al.. Journal of Clinical Investigation 94:797-807, 1994; and Inaba et al.. Journal of Experimental Medicine 172:631-640, 1990.

Assays for lymphocyte survival/apoptosis (which will identify, among others, proteins that prevent apoptosis after superantigen induction and proteins that regulate lymphocyte homeostasis) include, without limitation, those described in: Darzynkiewicz et al.. Cytometry 13:795-808, 1992; Gorczyca et al.. Leukemia 7:659-670, 1993; Gorczyca et al.. Cancer Research 53:1945-1951, 1993; Itoh et al.. Cell 66:233-243, 1991; Zacharchuk, Journal of Immunology 145:4037-4045, 1990; Zamai et al.. Cytometry 14:891-897, 1993; Gorczyca et al.. International Journal of Oncology 1:639-648, 1992.

Assays for proteins that influence early steps of T-cell commitment and development include, without limitation, those described in: Antica et al.. Blood 84:111-117, 1994; Fine et al.. Cellular Immunology 155:111-122, 1994; Galy et al.. Blood 85:2770-2778, 1995; Toki et al.. Proc. Nat. Acad Sci. USA 88:7548-7551, 1991.

5.10.8 ACTIVIN/INHIBIN ACTIVITY

A polypeptide of the present invention may also exhibit activin- or inhibin-related activities. A polynucleotide of the invention may encode a polypeptide exhibiting such characteristics. Inhibins are characterized by their ability to inhibit the release of follicle stimulating hormone (FSH), while activins and are characterized by their ability to stimulate the release of follicle stimulating hormone (FSH). Thus, a polypeptide of the present invention, alone or in heterodimers with a member of the inhibin family, may be useful as a contraceptive based on the ability of inhibins to decrease fertility in female mammals and decrease spermatogenesis in male mammals. Administration of sufficient amounts of other inhibins can induce infertility in these mammals. Alternatively, the polypeptide of the invention, as a homodimer or as a heterodimer with other protein subunits of the inhibin group, may be useful as a fertility inducing therapeutic, based upon the ability of activin molecules in stimulating FSH release from cells of the anterior pituitary. See, for example, U.S. Pat. No. 4,798,885. A polypeptide of the invention may also be useful for advancement of the onset of fertility in sexually immature mammals, so as to increase the lifetime reproductive performance of domestic animals such as, but not limited to, cows, sheep and pigs.

The activity of a polypeptide of the invention may, among other means, be measured by the following methods.

Assays for activin/inhibin activity include, without limitation, those described in: Vale et al.. Endocrinology 91:562-572, 1972; Ling et al.. Nature 321:779-782, 1986; Vale et al.. Nature 321:776-779, 1986; Mason et al.. Nature 318:659-663, 1985; Forage et al.. Proc. Natl. Acad. Sci. USA 83:3091-3095, 1986.

5.10.9 CHEMOTACTIC/CHEMOKINETIC ACTIVITY

A polypeptide of the present invention may be involved in chemotactic or chemokinetic activity for mammalian cells, including, for example, monocytes, fibroblasts, neutrophils, T-cells, mast cells, eosinophils, epithelial and/or endothelial cells. A polynucleotide of the invention can encode a polypeptide exhibiting such attributes. Chemotactic and chemokinetic receptor activation can be used to mobilize or attract a desired cell population to a desired site of action. Chemotactic or chemokinetic compositions (e.g. proteins, antibodies, binding partners, or modulators of the invention) provide particular advantages in treatment of wounds and other trauma to tissues, as well as in treatment of localized infections. For example, attraction of lymphocytes, monocytes or neutrophils to tumors or sites of infection may result in improved immune responses against the tumor or infecting agent.

A protein or peptide has chemotactic activity for a particular cell population if it can stimulate, directly or indirectly, the directed orientation or movement of such cell population. Preferably, the protein or peptide has the ability to directly stimulate directed movement of cells. Whether a particular protein has chemotactic activity for a population of cells can be readily determined by employing such protein or peptide in any known assay for cell chemotaxis.

Therapeutic compositions of the invention can be used in the following:

Assays for chemotactic activity (which will identify proteins that induce or prevent chemotaxis) consist of assays that measure the ability of a protein to induce the migration of cells across a membrane as well as the ability of a protein to induce the adhesion of one cell population to another cell population. Suitable assays for movement and adhesion include, without limitation, those described in: Current Protocols in Immunology, Ed by J. E. Coligan, A. M. Kruisbeek, D. H. Marguiles, E. M. Shevach, W. Strober, Pub. Greene Publishing Associates and Wiley-Interscience (Chapter 6.12, Measurement of alpha and beta Chemokines 6.12.1-6.12.28; Taub et. al., J. Clin. Invest. 95:1370-1376, 1995; Lind et. al., APMIS 103:140-146, 1995; Muller et al Eur. J. Immunol. 25:1744-1748; Gruber et. al., J. of Immunol. 152:5860-5867, 1994; Johnston et. al., J. of Immunol. 153:1762-1768, 1994.

5.10.10 HEMOSTATIC AND THROMBOLYTIC ACTIVITY

A polypeptide of the invention may also be involved in hemostatis or thrombolysis or thrombosis. A polynucleotide of the invention can encode a polypeptide exhibiting such attributes. Compositions may be useful in treatment of various coagulation disorders (including hereditary disorders, such as hemophilias) or to enhance coagulation and other hemostatic events in treating wounds resulting from trauma, surgery or other causes. A composition of the invention may also be useful for dissolving or inhibiting formation of thromboses and for treatment and prevention of conditions resulting therefrom (such as, for example, infarction of cardiac and central nervous system vessels (*e.g.*, stroke).

Therapeutic compositions of the invention can be used in the following:

Assay for hemostatic and thrombolytic activity include, without limitation, those described in: Linet et al.. J. Clin. Pharmacol. 26:131-140, 1986; Burdick et al.. Thrombosis Res. 45:413-419, 1987; Humphrey et al.. Fibrinolysis 5:71-79 (1991); Schaub, Prostaglandins 35:467-474, 1988.

5.10.11 CANCER DIAGNOSIS AND THERAPY

Polypeptides of the invention may be involved in cancer cell generation, proliferation or metastasis. Detection of the presence or amount of polynucleotides or polypeptides of the invention may be useful for the diagnosis and/or prognosis of one or more types of cancer. For example, the presence or increased expression of a polynucleotide/polypeptide of the invention may indicate a hereditary risk of cancer, a precancerous condition, or an ongoing malignancy. Conversely, a defect in the gene or absence of the polypeptide may be associated with a cancer condition. Identification of single nucleotide polymorphisms associated with cancer or a predisposition to cancer may also be useful for diagnosis or prognosis.

Cancer treatments promote tumor regression by inhibiting tumor cell proliferation, inhibiting angiogenesis (growth of new blood vessels that is necessary to support tumor growth) and/or prohibiting metastasis by reducing tumor cell motility or invasiveness. Therapeutic compositions of the invention may be effective in adult and pediatric oncology including in solid phase tumors/malignancies, locally advanced tumors, human soft tissue sarcomas, metastatic cancer, including lymphatic metastases,

blood cell malignancies including multiple myeloma, acute and chronic leukemias, and lymphomas, head and neck cancers including mouth cancer, larynx cancer and thyroid cancer, lung cancers including small cell carcinoma and non-small cell cancers, breast cancers including small cell carcinoma and ductal carcinoma, gastrointestinal cancers including esophageal cancer, stomach cancer, colon cancer, colorectal cancer and polyps associated with colorectal neoplasia, pancreatic cancers, liver cancer, urologic cancers including bladder cancer and prostate cancer, malignancies of the female genital tract including ovarian carcinoma, uterine (including endometrial) cancers, and solid tumor in the ovarian follicle, kidney cancers including renal cell carcinoma, brain cancers including intrinsic brain tumors, neuroblastoma, astrocytic brain tumors, gliomas, metastatic tumor cell invasion in the central nervous system, bone cancers including osteomas, skin cancers including malignant melanoma, tumor progression of human skin keratinocytes, squamous cell carcinoma, basal cell carcinoma, hemangiopericytoma and Karposi's sarcoma.

Polypeptides, polynucleotides, or modulators of polypeptides of the invention (including inhibitors and stimulators of the biological activity of the polypeptide of the invention) may be administered to treat cancer. Therapeutic compositions can be administered in therapeutically effective dosages alone or in combination with adjuvant cancer therapy such as surgery, chemotherapy, radiotherapy, thermotherapy, and laser therapy, and may provide a beneficial effect, e.g. reducing tumor size, slowing rate of tumor growth, inhibiting metastasis, or otherwise improving overall clinical condition, without necessarily eradicating the cancer.

The composition can also be administered in therapeutically effective amounts as a portion of an anti-cancer cocktail. An anti-cancer cocktail is a mixture of the polypeptide or modulator of the invention with one or more anti-cancer drugs in addition to a pharmaceutically acceptable carrier for delivery. The use of anti-cancer cocktails as a cancer treatment is routine. Anti-cancer drugs that are well known in the art and can be used as a treatment in combination with the polypeptide or modulator of the invention include: Actinomycin D, Aminoglutethimide, Asparaginase, Bleomycin, Busulfan, Carboplatin, Carmustine, Chlorambucil, Cisplatin (cis-DDP), Cyclophosphamide, Cytarabine HCl (Cytosine arabinoside), Dacarbazine, Dactinomycin, Daunorubicin HCl,

Doxorubicin HCl, Estramustine phosphate sodium, Etoposide (V16-213), Floxuridine, 5-Fluorouracil (5-Fu), Flutamide, Hydroxyurea (hydroxycarbamide), Ifosfamide, Interferon Alpha-2a, Interferon Alpha-2b, Leuprolide acetate (LHRH-releasing factor analog), Lomustine, Mechlorethamine HCl (nitrogen mustard), Melphalan, Mercaptopurine, Mesna, Methotrexate (MTX), Mitomycin, Mitoxantrone HCl, Octreotide, Plicamycin, Procarbazine HCl, Streptozocin, Tamoxifen citrate, Thioguanine, Thiotepa, Vinblastine sulfate, Vincristine sulfate, Amsacrine, Azacitidine, Hexamethylmelamine, Interleukin-2, Mitoguazone, Pentostatin, Semustine, Teniposide, and Vindesine sulfate.

In addition, therapeutic compositions of the invention may be used for prophylactic treatment of cancer. There are hereditary conditions and/or environmental situations (e.g. exposure to carcinogens) known in the art that predispose an individual to developing cancers. Under these circumstances, it may be beneficial to treat these individuals with therapeutically effective doses of the polypeptide of the invention to reduce the risk of developing cancers.

In vitro models can be used to determine the effective doses of the polypeptide of the invention as a potential cancer treatment. These *in vitro* models include proliferation assays of cultured tumor cells, growth of cultured tumor cells in soft agar (see Freshney, (1987) Culture of Animal Cells: A Manual of Basic Technique, Wily-Liss, New York, NY Ch 18 and Ch 21), tumor systems in nude mice as described in Giovanella et al.. J. Natl. Can. Inst., 52: 921-30 (1974), mobility and invasive potential of tumor cells in Boyden Chamber assays as described in Pilkington et al.. Anticancer Res., 17: 4107-9 (1997), and angiogenesis assays such as induction of vascularization of the chick chorioallantoic membrane or induction of vascular endothelial cell migration as described in Ribatta et al.. Intl. J. Dev. Biol., 40: 1189-97 (1999) and Li et al.. Clin. Exp. Metastasis, 17:423-9 (1999), respectively. Suitable tumor cells lines are available, e.g. from American Type Tissue Culture Collection catalogs.

5.10.12 RECEPTOR/LIGAND ACTIVITY

A polypeptide of the present invention may also demonstrate activity as receptor, receptor ligand or inhibitor or agonist of receptor/ligand interactions. A polynucleotide of the invention can encode a polypeptide exhibiting such characteristics. Examples of

such receptors and ligands include, without limitation, cytokine receptors and their ligands, receptor kinases and their ligands, receptor phosphatases and their ligands, receptors involved in cell-cell interactions and their ligands (including without limitation, cellular adhesion molecules (such as selectins, integrins and their ligands) and receptor/ligand pairs involved in antigen presentation, antigen recognition and development of cellular and humoral immune responses. Receptors and ligands are also useful for screening of potential peptide or small molecule inhibitors of the relevant receptor/ligand interaction. A protein of the present invention (including, without limitation, fragments of receptors and ligands) may themselves be useful as inhibitors of receptor/ligand interactions.

The activity of a polypeptide of the invention may, among other means, be measured by the following methods:

Suitable assays for receptor-ligand activity include without limitation those described in: Current Protocols in Immunology, Ed by J. E. Coligan, A. M. Kruisbeek, D. H. Margulies, E. M. Shevach, W. Strober, Pub. Greene Publishing Associates and Wiley- Interscience (Chapter 7.28, Measurement of Cellular Adhesion under static conditions 7.28.1- 7.28.22), Takai et al.. Proc. Natl. Acad. Sci. USA 84:6864-6868, 1987; Bierer et al.. J. Exp. Med. 168:1145-1156, 1988; Rosenstein et al.. J. Exp. Med. 169:149-160 1989; Stoltenborg et al.. J. Immunol. Methods 175:59-68, 1994; Stitt et al.. Cell 80:661-670, 1995.

By way of example, the polypeptides of the invention may be used as a receptor for a ligand(s) thereby transmitting the biological activity of that ligand(s). Ligands may be identified through binding assays, affinity chromatography, dihybrid screening assays, BIAcore assays, gel overlay assays, or other methods known in the art.

Studies characterizing drugs or proteins as agonist or antagonist or partial agonists or a partial antagonist require the use of other proteins as competing ligands. The polypeptides of the present invention or ligand(s) thereof may be labeled by being coupled to radioisotopes, colorimetric molecules or a toxin molecules by conventional methods. ("Guide to Protein Purification" Murray P. Deutscher (ed) Methods in Enzymology Vol. 182 (1990) Academic Press, Inc. San Diego). Examples of radioisotopes include, but are not limited to, tritium and carbon-14 . Examples of

colorimetric molecules include, but are not limited to, fluorescent molecules such as fluorescamine, or rhodamine or other colorimetric molecules. Examples of toxins include, but are not limited, to ricin.

5.10.13 DRUG SCREENING

This invention is particularly useful for screening chemical compounds by using the novel polypeptides or binding fragments thereof in any of a variety of drug screening techniques. The polypeptides or fragments employed in such a test may either be free in solution, affixed to a solid support, borne on a cell surface or located intracellularly. One method of drug screening utilizes eukaryotic or prokaryotic host cells which are stably transformed with recombinant nucleic acids expressing the polypeptide or a fragment thereof. Drugs are screened against such transformed cells in competitive binding assays. Such cells, either in viable or fixed form, can be used for standard binding assays. One may measure, for example, the formation of complexes between polypeptides of the invention or fragments and the agent being tested or examine the diminution in complex formation between the novel polypeptides and an appropriate cell line, which are well known in the art.

Sources for test compounds that may be screened for ability to bind to or modulate (*i.e.,* increase or decrease) the activity of polypeptides of the invention include (1) inorganic and organic chemical libraries, (2) natural product libraries, and (3) combinatorial libraries comprised of either random or mimetic peptides, oligonucleotides or organic molecules.

Chemical libraries may be readily synthesized or purchased from a number of commercial sources, and may include structural analogs of known compounds or compounds that are identified as "hits" or "leads" via natural product screening.

The sources of natural product libraries are microorganisms (including bacteria and fungi), animals, plants or other vegetation, or marine organisms, and libraries of mixtures for screening may be created by: (1) fermentation and extraction of broths from soil, plant or marine microorganisms or (2) extraction of the organisms themselves. Natural product libraries include polyketides, non-ribosomal peptides, and (non-naturally occurring) variants thereof. For a review, see *Science 282*:63-68 (1998).

Combinatorial libraries are composed of large numbers of peptides, oligonucleotides or organic compounds and can be readily prepared by traditional automated synthesis methods, PCR, cloning or proprietary synthetic methods. Of particular interest are peptide and oligonucleotide combinatorial libraries. Still other libraries of interest include peptide, protein, peptidomimetic, multiparallel synthetic collection, recombinatorial, and polypeptide libraries. For a review of combinatorial chemistry and libraries created therefrom, see Myers, *Curr. Opin. Biotechnol.* 8:701-707 (1997). For reviews and examples of peptidomimetic libraries, see Al-Obeidi et al.. *Mol. Biotechnol*, 9(3):205-23 (1998); Hruby et al.. *Curr Opin Chem Biol*, 1(1):114-19 (1997); Dorner et al.. *Bioorg Med Chem*, 4(5):709-15 (1996) (alkylated dipeptides).

Identification of modulators through use of the various libraries described herein permits modification of the candidate "hit" (or "lead") to optimize the capacity of the "hit" to bind a polypeptide of the invention. The molecules identified in the binding assay are then tested for antagonist or agonist activity in *in vivo* tissue culture or animal models that are well known in the art. In brief, the molecules are titrated into a plurality of cell cultures or animals and then tested for either cell/animal death or prolonged survival of the animal/cells.

The binding molecules thus identified may be complexed with toxins, *e.g.*, ricin or cholera, or with other compounds that are toxic to cells such as radioisotopes. The toxin-binding molecule complex is then targeted to a tumor or other cell by the specificity of the binding molecule for a polypeptide of the invention. Alternatively, the binding molecules may be complexed with imaging agents for targeting and imaging purposes.

5.10.14 ASSAY FOR RECEPTOR ACTIVITY

The invention also provides methods to detect specific binding of a polypeptide e.g. a ligand or a receptor. The art provides numerous assays particularly useful for identifying previously unknown binding partners for receptor polypeptides of the invention. For example, expression cloning using mammalian or bacterial cells, or dihybrid screening assays can be used to identify polynucleotides encoding binding partners. As another example, affinity chromatography with the appropriate immobilized polypeptide of the invention can be used to isolate polypeptides that recognize and bind

polypeptides of the invention. There are a number of different libraries used for the identification of compounds, and in particular small molecules, that modulate (*i.e.*, increase or decrease) biological activity of a polypeptide of the invention. Ligands for receptor polypeptides of the invention can also be identified by adding exogenous ligands, or cocktails of ligands to two cells populations that are genetically identical except for the expression of the receptor of the invention: one cell population expresses the receptor of the invention whereas the other does not. The response of the two cell populations to the addition of ligands(s) are then compared. Alternatively, an expression library can be co-expressed with the polypeptide of the invention in cells and assayed for an autocrine response to identify potential ligand(s). As still another example, BIAcore assays, gel overlay assays, or other methods known in the art can be used to identify binding partner polypeptides, including, (1) organic and inorganic chemical libraries, (2) natural product libraries, and (3) combinatorial libraries comprised of random peptides, oligonucleotides or organic molecules.

The role of downstream intracellular signaling molecules in the signaling cascade of the polypeptide of the invention can be determined. For example, a chimeric protein in which the cytoplasmic domain of the polypeptide of the invention is fused to the extracellular portion of a protein, whose ligand has been identified, is produced in a host cell. The cell is then incubated with the ligand specific for the extracellular portion of the chimeric protein, thereby activating the chimeric receptor. Known downstream proteins involved in intracellular signaling can then be assayed for expected modifications i.e. phosphorylation. Other methods known to those in the art can also be used to identify signaling molecules involved in receptor activity.

5.10.15 ANTI-INFLAMMATORY ACTIVITY

Compositions of the present invention may also exhibit anti-inflammatory activity. The anti-inflammatory activity may be achieved by providing a stimulus to cells involved in the inflammatory response, by inhibiting or promoting cell-cell interactions (such as, for example, cell adhesion), by inhibiting or promoting chemotaxis of cells involved in the inflammatory process, inhibiting or promoting cell extravasation, or by stimulating or suppressing production of other factors which more directly inhibit or

promote an inflammatory response. Compositions with such activities can be used to treat inflammatory conditions including chronic or acute conditions), including without limitation intimation associated with infection (such as septic shock, sepsis or systemic inflammatory response syndrome (SIRS)), ischemia-reperfusion injury, endotoxin lethality, arthritis, complement-mediated hyperacute rejection, nephritis, cytokine or chemokine-induced lung injury, inflammatory bowel disease, Crohn's disease or resulting from over production of cytokines such as TNF or IL-1. Compositions of the invention may also be useful to treat anaphylaxis and hypersensitivity to an antigenic substance or material. Compositions of this invention may be utilized to prevent or treat conditions such as, but not limited to, sepsis, acute pancreatitis, endotoxin shock, cytokine induced shock, rheumatoid arthritis, chronic inflammatory arthritis, pancreatic cell damage from diabetes mellitus type 1, graft versus host disease, inflammatory bowel disease, inflamation associated with pulmonary disease, other autoimmune disease or inflammatory disease, an antiproliferative agent such as for acute or chronic mylegenous leukemia or in the prevention of premature labor secondary to intrauterine infections.

5.10.16 LEUKEMIAS

Leukemias and related disorders may be treated or prevented by administration of a therapeutic that promotes or inhibits function of the polynucleotides and/or polypeptides of the invention. Such leukemias and related disorders include but are not limited to acute leukemia, acute lymphocytic leukemia, acute myelocytic leukemia, myeloblastic, promyelocytic, myelomonocytic, monocytic, erythroleukemia, chronic leukemia, chronic myelocytic (granulocytic) leukemia and chronic lymphocytic leukemia (for a review of such disorders, see Fishman et al.. 1985, Medicine, 2d Ed., J.B. Lippincott Co., Philadelphia).

5.10.17 NERVOUS SYSTEM DISORDERS

Nervous system disorders, involving cell types which can be tested for efficacy of intervention with compounds that modulate the activity of the polynucleotides and/or polypeptides of the invention, and which can be treated upon thus observing an indication of therapeutic utility, include but are not limited to nervous system injuries, and diseases or disorders which result in either a disconnection of axons, a diminution or degeneration of neurons, or demyelination. Nervous system lesions which may be treated in a patient (including human and non-human mammalian patients) according to the invention include but are not limited to the following lesions of either the central (including spinal cord, brain) or peripheral nervous systems:

(i) traumatic lesions, including lesions caused by physical injury or associated with surgery, for example, lesions which sever a portion of the nervous system, or compression injuries;

(ii) ischemic lesions, in which a lack of oxygen in a portion of the nervous system results in neuronal injury or death, including cerebral infarction or ischemia, or spinal cord infarction or ischemia;

(iii) infectious lesions, in which a portion of the nervous system is destroyed or injured as a result of infection, for example, by an abscess or associated with infection by human immunodeficiency virus, herpes zoster, or herpes simplex virus or with Lyme disease, tuberculosis, syphilis;

(iv) degenerative lesions, in which a portion of the nervous system is destroyed or injured as a result of a degenerative process including but not limited to degeneration associated with Parkinson's disease, Alzheimer's disease, Huntington's chorea, or amyotrophic lateral sclerosis;

(v) lesions associated with nutritional diseases or disorders, in which a portion of the nervous system is destroyed or injured by a nutritional disorder or disorder of metabolism including but not limited to, vitamin B12 deficiency, folic acid deficiency, Wernicke disease, tobacco-alcohol amblyopia, Marchiafava-Bignami disease (primary degeneration of the corpus callosum), and alcoholic cerebellar degeneration;

(vi) neurological lesions associated with systemic diseases including but not limited to diabetes (diabetic neuropathy, Bell's palsy), systemic lupus erythematosus, carcinoma, or sarcoidosis;

(vii) lesions caused by toxic substances including alcohol, lead, or particular neurotoxins; and

(viii) demyelinated lesions in which a portion of the nervous system is destroyed or injured by a demyelinating disease including but not limited to multiple sclerosis, human immunodeficiency virus-associated myelopathy, transverse myelopathy or various etiologies, progressive multifocal leukoencephalopathy, and central pontine myelinolysis.

Therapeutics which are useful according to the invention for treatment of a nervous system disorder may be selected by testing for biological activity in promoting the survival or differentiation of neurons. For example, and not by way of limitation, therapeutics which elicit any of the following effects may be useful according to the invention:

(i) increased survival time of neurons in culture;

(ii) increased sprouting of neurons in culture or *in vivo*;

(iii) increased production of a neuron-associated molecule in culture or *in vivo*, *e.g.*, choline acetyltransferase or acetylcholinesterase with respect to motor neurons; or

(iv) decreased symptoms of neuron dysfunction *in vivo*.

Such effects may be measured by any method known in the art. In preferred, non-limiting embodiments, increased survival of neurons may be measured by the method set forth in Arakawa et. al., (1990, J. Neurosci. 10:3507-3515); increased sprouting of neurons may be detected by methods set forth in Pestronk et. al., (1980, Exp. Neurol. 70:65-82) or Brown et. al., (1981, Ann. Rev. Neurosci. 4:17-42); increased production of neuron-associated molecules may be measured by bioassay, enzymatic assay, antibody binding, Northern blot assay, *etc.*, depending on the molecule to be measured; and motor neuron dysfunction may be measured by assessing the physical manifestation of motor neuron disorder, *e.g.*, weakness, motor neuron conduction velocity, or functional disability.

In specific embodiments, motor neuron disorders that may be treated according to the invention include but are not limited to disorders such as infarction, infection, exposure to toxin, trauma, surgical damage, degenerative disease or malignancy that may affect motor neurons as well as other components of the nervous system, as well as disorders that selectively affect neurons such as amyotrophic lateral sclerosis, and including but not limited to progressive spinal muscular atrophy, progressive bulbar palsy, primary lateral sclerosis, infantile and juvenile muscular atrophy, progressive bulbar paralysis of childhood (Fazio-Londe syndrome), poliomyelitis and the post polio syndrome, and Hereditary Motorsensory Neuropathy (Charcot-Marie-Tooth Disease).

5.10.18 OTHER ACTIVITIES

A polypeptide of the invention may also exhibit one or more of the following additional activities or effects: inhibiting the growth, infection or function of, or killing, infectious agents, including, without limitation, bacteria, viruses, fungi and other parasites; effecting (suppressing or enhancing) bodily characteristics, including, without limitation, height, weight, hair color, eye color, skin, fat to lean ratio or other tissue pigmentation, or organ or body part size or shape (such as, for example, breast augmentation or diminution, change in bone form or shape); effecting biorhythms or circadian cycles or rhythms; effecting the fertility of male or female subjects; effecting the metabolism, catabolism, anabolism, processing, utilization, storage or elimination of dietary fat, lipid, protein, carbohydrate, vitamins, minerals, co-factors or other nutritional factors or component(s); effecting behavioral characteristics, including, without limitation, appetite, libido, stress, cognition (including cognitive disorders), depression (including depressive disorders) and violent behaviors; providing analgesic effects or other pain reducing effects; promoting differentiation and growth of embryonic stem cells in lineages other than hematopoietic lineages; hormonal or endocrine activity; in the case of enzymes, correcting deficiencies of the enzyme and treating deficiency-related diseases; treatment of hyperproliferative disorders (such as, for example, psoriasis); immunoglobulin-like activity (such as, for example, the ability to bind antigens or complement); and the ability to act as an antigen in a vaccine composition to raise an

immune response against such protein or another material or entity which is cross-reactive with such protein.

5.10.19 IDENTIFICATION OF POLYMORPHISMS

The demonstration of polymorphisms makes possible the identification of such polymorphisms in human subjects and the pharmacogenetic use of this information for diagnosis and treatment. Such polymorphisms may be associated with, *e.g.*, differential predisposition or susceptibility to various disease states (such as disorders involving inflammation or immune response) or a differential response to drug administration, and this genetic information can be used to tailor preventive or therapeutic treatment appropriately. For example, the existence of a polymorphism associated with a predisposition to inflammation or autoimmune disease makes possible the diagnosis of this condition in humans by identifying the presence of the polymorphism.

Polymorphisms can be identified in a variety of ways known in the art which all generally involve obtaining a sample from a patient, analyzing DNA from the sample, optionally involving isolation or amplification of the DNA, and identifying the presence of the polymorphism in the DNA. For example, PCR may be used to amplify an appropriate fragment of genomic DNA which may then be sequenced. Alternatively, the DNA may be subjected to allele-specific oligonucleotide hybridization (in which appropriate oligonucleotides are hybridized to the DNA under conditions permitting detection of a single base mismatch) or to a single nucleotide extension assay (in which an oligonucleotide that hybridizes immediately adjacent to the position of the polymorphism is extended with one or more labeled nucleotides). In addition, traditional restriction fragment length polymorphism analysis (using restriction enzymes that provide differential digestion of the genomic DNA depending on the presence or absence of the polymorphism) may be performed. Arrays with nucleotide sequences of the present invention can be used to detect polymorphisms. The array can comprise modified nucleotide sequences of the present invention in order to detect the nucleotide sequences of the present invention. In the alternative, any one of the nucleotide sequences of the present invention can be placed on the array to detect changes from those sequences.

Alternatively a polymorphism resulting in a change in the amino acid sequence could also be detected by detecting a corresponding change in amino acid sequence of the protein, *e.g.*, by an antibody specific to the variant sequence.

5.10.20 ARTHRITIS AND INFLAMMATION

The immunosuppressive effects of the compositions of the invention against rheumatoid arthritis is determined in an experimental animal model system. The experimental model system is adjuvant induced arthritis in rats, and the protocol is described by J. Holoshitz, et at., 1983, Science, 219:56, or by B. Waksman et al.. 1963, Int. Arch. Allergy Appl. Immunol., 23:129. Induction of the disease can be caused by a single injection, generally intradermally, of a suspension of killed Mycobacterium tuberculosis in complete Freund's adjuvant (CFA). The route of injection can vary, but rats may be injected at the base of the tail with an adjuvant mixture. The polypeptide is administered in phosphate buffered solution (PBS) at a dose of about 1-5 mg/kg. The control consists of administering PBS only.

The procedure for testing the effects of the test compound would consist of intradermally injecting killed Mycobacterium tuberculosis in CFA followed by immediately administering the test compound and subsequent treatment every other day until day 24. At 14, 15, 18, 20, 22, and 24 days after injection of Mycobacterium CFA, an overall arthritis score may be obtained as described by J. Holoskitz above. An analysis of the data would reveal that the test compound would have a dramatic affect on the swelling of the joints as measured by a decrease of the arthritis score.

5.11 THERAPEUTIC METHODS

The compositions (including polypeptide fragments, analogs, variants and antibodies or other binding partners or modulators including antisense polynucleotides) of the invention have numerous applications in a variety of therapeutic methods. Examples of therapeutic applications include, but are not limited to, those exemplified herein.

5.11.1 EXAMPLE

One embodiment of the invention is the administration of an effective amount of the polypeptides or other composition of the invention to individuals affected by a disease or disorder that can be modulated by regulating the peptides of the invention. While the mode of administration is not particularly important, parenteral administration is preferred. An exemplary mode of administration is to deliver an intravenous bolus. The dosage of the polypeptides or other composition of the invention will normally be determined by the prescribing physician. It is to be expected that the dosage will vary according to the age, weight, condition and response of the individual patient. Typically, the amount of polypeptide administered per dose will be in the range of about 0.01μg/kg to 100 mg/kg of body weight, with the preferred dose being about 0.1μg/kg to 10 mg/kg of patient body weight. For parenteral administration, polypeptides of the invention will be formulated in an injectable form combined with a pharmaceutically acceptable parenteral vehicle. Such vehicles are well known in the art and examples include water, saline, Ringer's solution, dextrose solution, and solutions consisting of small amounts of the human serum albumin. The vehicle may contain minor amounts of additives that maintain the isotonicity and stability of the polypeptide or other active ingredient. The preparation of such solutions is within the skill of the art.

5.12 PHARMACEUTICAL FORMULATIONS AND ROUTES OF ADMINISTRATION

A protein or other composition of the present invention (from whatever source derived, including without limitation from recombinant and non-recombinant sources and including antibodies and other binding partners of the polypeptides of the invention) may be administered to a patient in need, by itself, or in pharmaceutical compositions where it is mixed with suitable carriers or excipient(s) at doses to treat or ameliorate a variety of disorders. Such a composition may optionally contain (in addition to protein or other active ingredient and a carrier) diluents, fillers, salts, buffers, stabilizers, solubilizers, and other materials well known in the art. The term "pharmaceutically acceptable" means a non-toxic material that does not interfere with the effectiveness of the biological activity of the active ingredient(s). The characteristics of the carrier will depend on the route of

administration. The pharmaceutical composition of the invention may also contain cytokines, lymphokines, or other hematopoietic factors such as M-CSF, GM-CSF, TNF, IL-1, IL-2, IL-3, IL-4, IL-5, IL-6, IL-7, IL-8, IL-9, IL-10, IL-11, IL-12, IL-13, IL-14, IL-15, IFN, TNF0, TNF1, TNF2, G-CSF, Meg-CSF, thrombopoietin, stem cell factor, and erythropoietin. In further compositions, proteins of the invention may be combined with other agents beneficial to the treatment of the disease or disorder in question. These agents include various growth factors such as epidermal growth factor (EGF), platelet-derived growth factor (PDGF), transforming growth factors (TGF-α and TGF-β), insulin-like growth factor (IGF), as well as cytokines described herein.

The pharmaceutical composition may further contain other agents which either enhance the activity of the protein or other active ingredient or complement its activity or use in treatment. Such additional factors and/or agents may be included in the pharmaceutical composition to produce a synergistic effect with protein or other active ingredient of the invention, or to minimize side effects. Conversely, protein or other active ingredient of the present invention may be included in formulations of the particular clotting factor, cytokine, lymphokine, other hematopoietic factor, thrombolytic or anti-thrombotic factor, or anti- inflammatory agent to minimize side effects of the clotting factor, cytokine, lymphokine, other hematopoietic factor, thrombolytic or anti-thrombotic factor, or anti-inflammatory agent (such as IL-1Ra, IL-1 Hy1, IL-1 Hy2, anti-TNF, corticosteroids, immunosuppressive agents). A protein of the present invention may be active in multimers (*e.g.*, heterodimers or homodimers) or complexes with itself or other proteins. As a result, pharmaceutical compositions of the invention may comprise a protein of the invention in such multimeric or complexed form.

As an alternative to being included in a pharmaceutical composition of the invention including a first protein, a second protein or a therapeutic agent may be concurrently administered with the first protein (*e.g.*, at the same time, or at differing times provided that therapeutic concentrations of the combination of agents is achieved at the treatment site). Techniques for formulation and administration of the compounds of the instant application may be found in "Remington's Pharmaceutical Sciences," Mack Publishing Co., Easton, PA, 18th edition. A therapeutically effective dose further refers to that amount of the compound sufficient to result in amelioration of symptoms, *e.g.*,

treatment, healing, prevention or amelioration of the relevant medical condition, or an increase in rate of treatment, healing, prevention or amelioration of such conditions. When applied to an individual active ingredient, administered alone, a therapeutically effective dose refers to that ingredient alone. When applied to a combination, a therapeutically effective dose refers to combined amounts of the active ingredients that result in the therapeutic effect, whether administered in combination, serially or simultaneously.

In practicing the method of treatment or use of the present invention, a therapeutically effective amount of protein or other active ingredient of the present invention is administered to a mammal having a condition to be treated. Protein or other active ingredient of the present invention may be administered in accordance with the method of the invention either alone or in combination with other therapies such as treatments employing cytokines, lymphokines or other hematopoietic factors. When co-administered with one or more cytokines, lymphokines or other hematopoietic factors, protein or other active ingredient of the present invention may be administered either simultaneously with the cytokine(s), lymphokine(s), other hematopoietic factor(s), thrombolytic or anti-thrombotic factors, or sequentially. If administered sequentially, the attending physician will decide on the appropriate sequence of administering protein or other active ingredient of the present invention in combination with cytokine(s), lymphokine(s), other hematopoietic factor(s), thrombolytic or anti-thrombotic factors.

5.12.1 ROUTES OF ADMINISTRATION

Suitable routes of administration may, for example, include oral, rectal, transmucosal, or intestinal administration; parenteral delivery, including intramuscular, subcutaneous, intramedullary injections, as well as intrathecal, direct intraventricular, intravenous, intraperitoneal, intranasal, or intraocular injections. Administration of protein or other active ingredient of the present invention used in the pharmaceutical composition or to practice the method of the present invention can be carried out in a variety of conventional ways, such as oral ingestion, inhalation, topical application or cutaneous, subcutaneous, intraperitoneal, parenteral or intravenous injection. Intravenous administration to the patient is preferred.

Alternately, one may administer the compound in a local rather than systemic manner, for example, via injection of the compound directly into a arthritic joints or in fibrotic tissue, often in a depot or sustained release formulation. In order to prevent the scarring process frequently occurring as complication of glaucoma surgery, the compounds may be administered topically, for example, as eye drops. Furthermore, one may administer the drug in a targeted drug delivery system, for example, in a liposome coated with a specific antibody, targeting, for example, arthritic or fibrotic tissue. The liposomes will be targeted to and taken up selectively by the afflicted tissue.

The polypeptides of the invention are administered by any route that delivers an effective dosage to the desired site of action. The determination of a suitable route of administration and an effective dosage for a particular indication is within the level of skill in the art. Preferably for wound treatment, one administers the therapeutic compound directly to the site. Suitable dosage ranges for the polypeptides of the invention can be extrapolated from these dosages or from similar studies in appropriate animal models. Dosages can then be adjusted as necessary by the clinician to provide maximal therapeutic benefit.

5.12.2 COMPOSITIONS/FORMULATIONS

Pharmaceutical compositions for use in accordance with the present invention thus may be formulated in a conventional manner using one or more physiologically acceptable carriers comprising excipients and auxiliaries which facilitate processing of the active compounds into preparations which can be used pharmaceutically. These pharmaceutical compositions may be manufactured in a manner that is itself known, *e.g.*, by means of conventional mixing, dissolving, granulating, dragee-making, levigating, emulsifying, encapsulating, entrapping or lyophilizing processes. Proper formulation is dependent upon the route of administration chosen. When a therapeutically effective amount of protein or other active ingredient of the present invention is administered orally, protein or other active ingredient of the present invention will be in the form of a tablet, capsule, powder, solution or elixir. When administered in tablet form, the pharmaceutical composition of the invention may additionally contain a solid carrier such as a gelatin or an adjuvant. The tablet, capsule, and powder contain from about 5 to 95%

protein or other active ingredient of the present invention, and preferably from about 25 to 90% protein or other active ingredient of the present invention. When administered in liquid form, a liquid carrier such as water, petroleum, oils of animal or plant origin such as peanut oil, mineral oil, soybean oil, or sesame oil, or synthetic oils may be added. The liquid form of the pharmaceutical composition may further contain physiological saline solution, dextrose or other saccharide solution, or glycols such as ethylene glycol, propylene glycol or polyethylene glycol. When administered in liquid form, the pharmaceutical composition contains from about 0.5 to 90% by weight of protein or other active ingredient of the present invention, and preferably from about 1 to 50% protein or other active ingredient of the present invention.

When a therapeutically effective amount of protein or other active ingredient of the present invention is administered by intravenous, cutaneous or subcutaneous injection, protein or other active ingredient of the present invention will be in the form of a pyrogen-free, parenterally acceptable aqueous solution. The preparation of such parenterally acceptable protein or other active ingredient solutions, having due regard to pH, isotonicity, stability, and the like, is within the skill in the art. A preferred pharmaceutical composition for intravenous, cutaneous, or subcutaneous injection should contain, in addition to protein or other active ingredient of the present invention, an isotonic vehicle such as sodium chloride injection, Ringer's injection, dextrose injection, dextrose and sodium chloride injection, lactated Ringer's injection, or other vehicle as known in the art. The pharmaceutical composition of the present invention may also contain stabilizers, preservatives, buffers, antioxidants, or other additives known to those of skill in the art. For injection, the agents of the invention may be formulated in aqueous solutions, preferably in physiologically compatible buffers such as Hanks's solution, Ringer's solution, or physiological saline buffer. For transmucosal administration, penetrants appropriate to the barrier to be permeated are used in the formulation. Such penetrants are generally known in the art.

For oral administration, the compounds can be formulated readily by combining the active compounds with pharmaceutically acceptable carriers well known in the art. Such carriers enable the compounds of the invention to be formulated as tablets, pills, dragees, capsules, liquids, gels, syrups, slurries, suspensions and the like, for oral

ingestion by a patient to be treated. Pharmaceutical preparations for oral use can be obtained from a solid excipient, optionally grinding a resulting mixture, and processing the mixture of granules, after adding suitable auxiliaries, if desired, to obtain tablets or dragee cores. Suitable excipients are, in particular, fillers such as sugars, including lactose, sucrose, mannitol, or sorbitol; cellulose preparations such as, for example, maize starch, wheat starch, rice starch, potato starch, gelatin, gum tragacanth, methyl cellulose, hydroxypropylmethyl-cellulose, sodium carboxymethylcellulose, and/or polyvinylpyrrolidone (PVP). If desired, disintegrating agents may be added, such as the cross-linked polyvinyl pyrrolidone, agar, or alginic acid or a salt thereof such as sodium alginate. Dragee cores are provided with suitable coatings. For this purpose, concentrated sugar solutions may be used, which may optionally contain gum arabic, talc, polyvinyl pyrrolidone, carbopol gel, polyethylene glycol, and/or titanium dioxide, lacquer solutions, and suitable organic solvents or solvent mixtures. Dyestuffs or pigments may be added to the tablets or dragee coatings for identification or to characterize different combinations of active compound doses.

Pharmaceutical preparations which can be used orally include push-fit capsules made of gelatin, as well as soft, sealed capsules made of gelatin and a plasticizer, such as glycerol or sorbitol. The push-fit capsules can contain the active ingredients in admixture with filler such as lactose, binders such as starches, and/or lubricants such as talc or magnesium stearate and, optionally, stabilizers. In soft capsules, the active compounds may be dissolved or suspended in suitable liquids, such as fatty oils, liquid paraffin, or liquid polyethylene glycols. In addition, stabilizers may be added. All formulations for oral administration should be in dosages suitable for such administration. For buccal administration, the compositions may take the form of tablets or lozenges formulated in conventional manner.

For administration by inhalation, the compounds for use according to the present invention are conveniently delivered in the form of an aerosol spray presentation from pressurized packs or a nebuliser, with the use of a suitable propellant, *e.g.*, dichlorodifluoromethane, trichlorofluoromethane, dichlorotetrafluoroethane, carbon dioxide or other suitable gas. In the case of a pressurized aerosol the dosage unit may be determined by providing a valve to deliver a metered amount. Capsules and cartridges

of, *e.g.*, gelatin for use in an inhaler or insufflator may be formulated containing a powder mix of the compound and a suitable powder base such as lactose or starch. The compounds may be formulated for parenteral administration by injection, *e.g.*, by bolus injection or continuous infusion. Formulations for injection may be presented in unit dosage form, *e.g.*, in ampules or in multi-dose containers, with an added preservative. The compositions may take such forms as suspensions, solutions or emulsions in oily or aqueous vehicles, and may contain formulatory agents such as suspending, stabilizing and/or dispersing agents.

Pharmaceutical formulations for parenteral administration include aqueous solutions of the active compounds in water-soluble form. Additionally, suspensions of the active compounds may be prepared as appropriate oily injection suspensions. Suitable lipophilic solvents or vehicles include fatty oils such as sesame oil, or synthetic fatty acid esters, such as ethyl oleate or triglycerides, or liposomes. Aqueous injection suspensions may contain substances which increase the viscosity of the suspension, such as sodium carboxymethyl cellulose, sorbitol, or dextran. Optionally, the suspension may also contain suitable stabilizers or agents which increase the solubility of the compounds to allow for the preparation of highly concentrated solutions. Alternatively, the active ingredient may be in powder form for constitution with a suitable vehicle, *e.g.*, sterile pyrogen-free water, before use.

The compounds may also be formulated in rectal compositions such as suppositories or retention enemas, *e.g.*, containing conventional suppository bases such as cocoa butter or other glycerides. In addition to the formulations described previously, the compounds may also be formulated as a depot preparation. Such long acting formulations may be administered by implantation (for example subcutaneously or intramuscularly) or by intramuscular injection. Thus, for example, the compounds may be formulated with suitable polymeric or hydrophobic materials (for example as an emulsion in an acceptable oil) or ion exchange resins, or as sparingly soluble derivatives, for example, as a sparingly soluble salt.

A pharmaceutical carrier for the hydrophobic compounds of the invention is a co-solvent system comprising benzyl alcohol, a nonpolar surfactant, a water-miscible organic polymer, and an aqueous phase. The co-solvent system may be the VPD

co-solvent system. VPD is a solution of 3% w/v benzyl alcohol, 8% w/v of the nonpolar surfactant polysorbate 80, and 65% w/v polyethylene glycol 300, made up to volume in absolute ethanol. The VPD co-solvent system (VPD:5W) consists of VPD diluted 1:1 with a 5% dextrose in water solution. This co-solvent system dissolves hydrophobic compounds well, and itself produces low toxicity upon systemic administration. Naturally, the proportions of a co-solvent system may be varied considerably without destroying its solubility and toxicity characteristics. Furthermore, the identity of the co-solvent components may be varied: for example, other low-toxicity nonpolar surfactants may be used instead of polysorbate 80; the fraction size of polyethylene glycol may be varied; other biocompatible polymers may replace polyethylene glycol, e.g. polyvinyl pyrrolidone; and other sugars or polysaccharides may substitute for dextrose. Alternatively, other delivery systems for hydrophobic pharmaceutical compounds may be employed. Liposomes and emulsions are well known examples of delivery vehicles or carriers for hydrophobic drugs. Certain organic solvents such as dimethylsulfoxide also may be employed, although usually at the cost of greater toxicity. Additionally, the compounds may be delivered using a sustained-release system, such as semipermeable matrices of solid hydrophobic polymers containing the therapeutic agent. Various types of sustained-release materials have been established and are well known by those skilled in the art. Sustained-release capsules may, depending on their chemical nature, release the compounds for a few weeks up to over 100 days. Depending on the chemical nature and the biological stability of the therapeutic reagent, additional strategies for protein or other active ingredient stabilization may be employed.

The pharmaceutical compositions also may comprise suitable solid or gel phase carriers or excipients. Examples of such carriers or excipients include but are not limited to calcium carbonate, calcium phosphate, various sugars, starches, cellulose derivatives, gelatin, and polymers such as polyethylene glycols. Many of the active ingredients of the invention may be provided as salts with pharmaceutically compatible counter ions. Such pharmaceutically acceptable base addition salts are those salts which retain the biological effectiveness and properties of the free acids and which are obtained by reaction with inorganic or organic bases such as sodium hydroxide, magnesium hydroxide, ammonia,

trialkylamine, dialkylamine, monoalkylamine, dibasic amino acids, sodium acetate, potassium benzoate, triethanol amine and the like.

The pharmaceutical composition of the invention may be in the form of a complex of the protein(s) or other active ingredient(s) of present invention along with protein or peptide antigens. The protein and/or peptide antigen will deliver a stimulatory signal to both B and T lymphocytes. B lymphocytes will respond to antigen through their surface immunoglobulin receptor. T lymphocytes will respond to antigen through the T cell receptor (TCR) following presentation of the antigen by MHC proteins. MHC and structurally related proteins including those encoded by class I and class II MHC genes on host cells will serve to present the peptide antigen(s) to T lymphocytes. The antigen components could also be supplied as purified MHC-peptide complexes alone or with co-stimulatory molecules that can directly signal T cells. Alternatively antibodies able to bind surface immunoglobulin and other molecules on B cells as well as antibodies able to bind the TCR and other molecules on T cells can be combined with the pharmaceutical composition of the invention.

The pharmaceutical composition of the invention may be in the form of a liposome in which protein of the present invention is combined, in addition to other pharmaceutically acceptable carriers, with amphipathic agents such as lipids which exist in aggregated form as micelles, insoluble monolayers, liquid crystals, or lamellar layers in aqueous solution. Suitable lipids for liposomal formulation include, without limitation, monoglycerides, diglycerides, sulfatides, lysolecithins, phospholipids, saponin, bile acids, and the like. Preparation of such liposomal formulations is within the level of skill in the art, as disclosed, for example, in U.S. Patent Nos. 4,235,871; 4,501,728; 4,837,028; and 4,737,323, all of which are incorporated herein by reference.

The amount of protein or other active ingredient of the present invention in the pharmaceutical composition of the present invention will depend upon the nature and severity of the condition being treated, and on the nature of prior treatments which the patient has undergone. Ultimately, the attending physician will decide the amount of protein or other active ingredient of the present invention with which to treat each individual patient. Initially, the attending physician will administer low doses of protein or other active ingredient of the present invention and observe the patient's response.

Larger doses of protein or other active ingredient of the present invention may be administered until the optimal therapeutic effect is obtained for the patient, and at that point the dosage is not increased further. It is contemplated that the various pharmaceutical compositions used to practice the method of the present invention should contain about 0.01 μg to about 100 mg (preferably about 0.1 μg to about 10 mg, more preferably about 0.1 μg to about 1 mg) of protein or other active ingredient of the present invention per kg body weight. For compositions of the present invention which are useful for bone, cartilage, tendon or ligament regeneration, the therapeutic method includes administering the composition topically, systematically, or locally as an implant or device. When administered, the therapeutic composition for use in this invention is, of course, in a pyrogen-free, physiologically acceptable form. Further, the composition may desirably be encapsulated or injected in a viscous form for delivery to the site of bone, cartilage or tissue damage. Topical administration may be suitable for wound healing and tissue repair. Therapeutically useful agents other than a protein or other active ingredient of the invention which may also optionally be included in the composition as described above, may alternatively or additionally, be administered simultaneously or sequentially with the composition in the methods of the invention. Preferably for bone and/or cartilage formation, the composition would include a matrix capable of delivering the protein-containing or other active ingredient-containing composition to the site of bone and/or cartilage damage, providing a structure for the developing bone and cartilage and optimally capable of being resorbed into the body. Such matrices may be formed of materials presently in use for other implanted medical applications.

The choice of matrix material is based on biocompatibility, biodegradability, mechanical properties, cosmetic appearance and interface properties. The particular application of the compositions will define the appropriate formulation. Potential matrices for the compositions may be biodegradable and chemically defined calcium sulfate, tricalcium phosphate, hydroxyapatite, polylactic acid, polyglycolic acid and polyanhydrides. Other potential materials are biodegradable and biologically well-defined, such as bone or dermal collagen. Further matrices are comprised of pure proteins or extracellular matrix components. Other potential matrices are nonbiodegradable and chemically defined, such as sintered hydroxyapatite, bioglass,

aluminates, or other ceramics. Matrices may be comprised of combinations of any of the above mentioned types of material, such as polylactic acid and hydroxyapatite or collagen and tricalcium phosphate. The bioceramics may be altered in composition, such as in calcium-aluminate-phosphate and processing to alter pore size, particle size, particle shape, and biodegradability. Presently preferred is a 50:50 (mole weight) copolymer of lactic acid and glycolic acid in the form of porous particles having diameters ranging from 150 to 800 microns. In some applications, it will be useful to utilize a sequestering agent, such as carboxymethyl cellulose or autologous blood clot, to prevent the protein compositions from disassociating from the matrix.

A preferred family of sequestering agents is cellulosic materials such as alkylcelluloses (including hydroxyalkylcelluloses), including methylcellulose, ethylcellulose, hydroxyethylcellulose, hydroxypropylcellulose, hydroxypropyl-methylcellulose, and carboxymethylcellulose, the most preferred being cationic salts of carboxymethylcellulose (CMC). Other preferred sequestering agents include hyaluronic acid, sodium alginate, poly(ethylene glycol), polyoxyethylene oxide, carboxyvinyl polymer and poly(vinyl alcohol). The amount of sequestering agent useful herein is 0.5-20 wt %, preferably 1-10 wt % based on total formulation weight, which represents the amount necessary to prevent desorption of the protein from the polymer matrix and to provide appropriate handling of the composition, yet not so much that the progenitor cells are prevented from infiltrating the matrix, thereby providing the protein the opportunity to assist the osteogenic activity of the progenitor cells. In further compositions, proteins or other active ingredients of the invention may be combined with other agents beneficial to the treatment of the bone and/or cartilage defect, wound, or tissue in question. These agents include various growth factors such as epidermal growth factor (EGF), platelet derived growth factor (PDGF), transforming growth factors (TGF-α and TGF-β), and insulin-like growth factor (IGF).

The therapeutic compositions are also presently valuable for veterinary applications. Particularly domestic animals and thoroughbred horses, in addition to humans, are desired patients for such treatment with proteins or other active ingredients of the present invention. The dosage regimen of a protein-containing pharmaceutical composition to be used in tissue regeneration will be determined by the attending

physician considering various factors which modify the action of the proteins, *e.g.*, amount of tissue weight desired to be formed, the site of damage, the condition of the damaged tissue, the size of a wound, type of damaged tissue (*e.g.*, bone), the patient's age, sex, and diet, the severity of any infection, time of administration and other clinical factors. The dosage may vary with the type of matrix used in the reconstitution and with inclusion of other proteins in the pharmaceutical composition. For example, the addition of other known growth factors, such as IGF I (insulin like growth factor I), to the final composition, may also effect the dosage. Progress can be monitored by periodic assessment of tissue/bone growth and/or repair, for example, X-rays, histomorphometric determinations and tetracycline labeling.

Polynucleotides of the present invention can also be used for gene therapy. Such polynucleotides can be introduced either *in vivo* or *ex vivo* into cells for expression in a mammalian subject. Polynucleotides of the invention may also be administered by other known methods for introduction of nucleic acid into a cell or organism (including, without limitation, in the form of viral vectors or naked DNA). Cells may also be cultured *ex vivo* in the presence of proteins of the present invention in order to proliferate or to produce a desired effect on or activity in such cells. Treated cells can then be introduced *in vivo* for therapeutic purposes.

5.12.3 EFFECTIVE DOSAGE

Pharmaceutical compositions suitable for use in the present invention include compositions wherein the active ingredients are contained in an effective amount to achieve its intended purpose. More specifically, a therapeutically effective amount means an amount effective to prevent development of or to alleviate the existing symptoms of the subject being treated. Determination of the effective amount is well within the capability of those skilled in the art, especially in light of the detailed disclosure provided herein. For any compound used in the method of the invention, the therapeutically effective dose can be estimated initially from appropriate *in vitro* assays. For example, a dose can be formulated in animal models to achieve a circulating concentration range that can be used to more accurately determine useful doses in humans. For example, a dose can be formulated in animal models to achieve a

circulating concentration range that includes the IC_{50} as determined in cell culture (*i.e.*, the concentration of the test compound which achieves a half-maximal inhibition of the protein's biological activity). Such information can be used to more accurately determine useful doses in humans.

A therapeutically effective dose refers to that amount of the compound that results in amelioration of symptoms or a prolongation of survival in a patient. Toxicity and therapeutic efficacy of such compounds can be determined by standard pharmaceutical procedures in cell cultures or experimental animals, *e.g.*, for determining the LD_{50} (the dose lethal to 50% of the population) and the ED_{50} (the dose therapeutically effective in 50% of the population). The dose ratio between toxic and therapeutic effects is the therapeutic index and it can be expressed as the ratio between LD_{50} and ED_{50}. Compounds which exhibit high therapeutic indices are preferred. The data obtained from these cell culture assays and animal studies can be used in formulating a range of dosage for use in human. The dosage of such compounds lies preferably within a range of circulating concentrations that include the ED_{50} with little or no toxicity. The dosage may vary within this range depending upon the dosage form employed and the route of administration utilized. The exact formulation, route of administration and dosage can be chosen by the individual physician in view of the patient's condition. See, *e.g.*, Fingl et al.. 1975, in "The Pharmacological Basis of Therapeutics", Ch. 1 p.1. Dosage amount and interval may be adjusted individually to provide plasma levels of the active moiety which are sufficient to maintain the desired effects, or minimal effective concentration (MEC). The MEC will vary for each compound but can be estimated from *in vitro* data. Dosages necessary to achieve the MEC will depend on individual characteristics and route of administration. However, HPLC assays or bioassays can be used to determine plasma concentrations.

Dosage intervals can also be determined using MEC value. Compounds should be administered using a regimen which maintains plasma levels above the MEC for 10-90% of the time, preferably between 30-90% and most preferably between 50-90%. In cases of local administration or selective uptake, the effective local concentration of the drug may not be related to plasma concentration.

An exemplary dosage regimen for polypeptides or other compositions of the invention will be in the range of about 0.01 μg/kg to 100 mg/kg of body weight daily, with the preferred dose being about 0.1 μg/kg to 25 mg/kg of patient body weight daily, varying in adults and children. Dosing may be once daily, or equivalent doses may be delivered at longer or shorter intervals.

The amount of composition administered will, of course, be dependent on the subject being treated, on the subject's age and weight, the severity of the affliction, the manner of administration and the judgment of the prescribing physician.

5.12.4 PACKAGING

The compositions may, if desired, be presented in a pack or dispenser device which may contain one or more unit dosage forms containing the active ingredient. The pack may, for example, comprise metal or plastic foil, such as a blister pack. The pack or dispenser device may be accompanied by instructions for administration. Compositions comprising a compound of the invention formulated in a compatible pharmaceutical carrier may also be prepared, placed in an appropriate container, and labeled for treatment of an indicated condition.

5.13 ANTIBODIES

Also included in the invention are antibodies to proteins, or fragments of proteins of the invention. The term "antibody" as used herein refers to immunoglobulin molecules and immunologically active portions of immunoglobulin (Ig) molecules, *i.e.*, molecules that contain an antigen binding site that specifically binds (immunoreacts with) an antigen. Such antibodies include, but are not limited to, polyclonal, monoclonal, chimeric, single chain, F_{ab}, $F_{ab'}$ and $F_{(ab')2}$ fragments, and an F_{ab} expression library. In general, an antibody molecule obtained from humans relates to any of the classes IgG, IgM, IgA, IgE and IgD, which differ from one another by the nature of the heavy chain present in the molecule. Certain classes have subclasses as well, such as IgG_1, IgG_2, and others. Furthermore, in humans, the light chain may be a kappa chain or a lambda chain. Reference herein to antibodies includes a reference to all such classes, subclasses and types of human antibody species.

An isolated related protein of the invention may be intended to serve as an antigen, or a portion or fragment thereof, and additionally can be used as an immunogen to generate antibodies that immunospecifically bind the antigen, using standard techniques for polyclonal and monoclonal antibody preparation. The full-length protein can be used or, alternatively, the invention provides antigenic peptide fragments of the antigen for use as immunogens. An antigenic peptide fragment comprises at least 6 amino acid residues of the amino acid sequence of the full length protein, such as an amino acid sequence encoded by 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502, and encompasses an epitope thereof such that an antibody raised against the peptide forms a specific immune complex with the full length protein or with any fragment that contains the epitope. Preferably, the antigenic peptide comprises at least 10 amino acid residues, or at least 15 amino acid residues, or at least 20 amino acid residues, or at least 30 amino acid residues. Preferred epitopes encompassed by the antigenic peptide are regions of the protein that are located on its surface; commonly these are hydrophilic regions.

In certain embodiments of the invention, at least one epitope encompassed by the antigenic peptide is a region of -related protein that is located on the surface of the protein, *e.g.*, a hydrophilic region. A hydrophobicity analysis of the human related protein sequence will indicate which regions of a related protein are particularly hydrophilic and, therefore, are likely to encode surface residues useful for targeting antibody production. As a means for targeting antibody production, hydropathy plots showing regions of hydrophilicity and hydrophobicity may be generated by any method well known in the art, including, for example, the Kyte Doolittle or the Hopp Woods methods, either with or without Fourier transformation. See, *e.g.*, Hopp and Woods, 1981, *Proc. Nat. Acad. Sci. USA* 78: 3824-3828; Kyte and Doolittle 1982, *J. Mol. Biol.* 157: 105-142, each of which is incorporated herein by reference in its entirety. Antibodies that are specific for one or more domains within an antigenic protein, or derivatives, fragments, analogs or homologs thereof, are also provided herein.

A protein of the invention, or a derivative, fragment, analog, homolog or ortholog thereof, may be utilized as an immunogen in the generation of antibodies that immunospecifically bind these protein components.

Various procedures known within the art may be used for the production of polyclonal or monoclonal antibodies directed against a protein of the invention, or against derivatives, fragments, analogs homologs or orthologs thereof (see, for example, Antibodies: A Laboratory Manual, Harlow E, and Lane D, 1988, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, incorporated herein by reference). Some of these antibodies are discussed below.

5.13.1 Polyclonal Antibodies

For the production of polyclonal antibodies, various suitable host animals (*e.g.*, rabbit, goat, mouse or other mammal) may be immunized by one or more injections with the native protein, a synthetic variant thereof, or a derivative of the foregoing. An appropriate immunogenic preparation can contain, for example, the naturally occurring immunogenic protein, a chemically synthesized polypeptide representing the immunogenic protein, or a recombinantly expressed immunogenic protein. Furthermore, the protein may be conjugated to a second protein known to be immunogenic in the mammal being immunized. Examples of such immunogenic proteins include but are not limited to keyhole limpet hemocyanin, serum albumin, bovine thyroglobulin, and soybean trypsin inhibitor. The preparation can further include an adjuvant. Various adjuvants used to increase the immunological response include, but are not limited to, Freund's (complete and incomplete), mineral gels (*e.g.*, aluminum hydroxide), surface active substances (*e.g.*, lysolecithin, pluronic polyols, polyanions, peptides, oil emulsions, dinitrophenol, etc.), adjuvants usable in humans such as Bacille Calmette-Guerin (BCG) and Corynebacterium parvum, or similar immunostimulatory agents. Additional examples of adjuvants which can be employed include MPL-TDM adjuvant (monophosphoryl Lipid A, synthetic trehalose dicorynomycolate).

The polyclonal antibody molecules directed against the immunogenic protein can be isolated from the mammal (*e.g.*, from the blood) and further purified by well known techniques, such as affinity chromatography using protein A or protein G, which provide

primarily the IgG fraction of immune serum. Subsequently, or alternatively, the specific antigen which is the target of the immunoglobulin sought, or an epitope thereof, may be immobilized on a column to purify the immune specific antibody by immunoaffinity chromatography. Purification of immunoglobulins is discussed, for example, by D. Wilkinson (The Scientist, published by The Scientist, Inc., Philadelphia PA, Vol. 14, No. 8 (April 17, 2000), pp. 25-28).

5.13.2 Monoclonal Antibodies

The term "monoclonal antibody" (MAb) or "monoclonal antibody composition", as used herein, refers to a population of antibody molecules that contain only one molecular species of antibody molecule consisting of a unique light chain gene product and a unique heavy chain gene product. In particular, the complementarity determining regions (CDRs) of the monoclonal antibody are identical in all the molecules of the population. MAbs thus contain an antigen binding site capable of immunoreacting with a particular epitope of the antigen characterized by a unique binding affinity for it.

Monoclonal antibodies can be prepared using hybridoma methods, such as those described by Kohler and Milstein, Nature, 256:495 (1975). In a hybridoma method, a mouse, hamster, or other appropriate host animal, is typically immunized with an immunizing agent to elicit lymphocytes that produce or are capable of producing antibodies that will specifically bind to the immunizing agent. Alternatively, the lymphocytes can be immunized *in vitro*. The immunizing agent will typically include the protein antigen, a fragment thereof or a fusion protein thereof. Generally, either peripheral blood lymphocytes are used if cells of human origin are desired, or spleen cells or lymph node cells are used if non-human mammalian sources are desired. The lymphocytes are then fused with an immortalized cell line using a suitable fusing agent, such as polyethylene glycol, to form a hybridoma cell (Goding, Monoclonal Antibodies: Principles and Practice, Academic Press, (1986) pp. 59-103). Immortalized cell lines are usually transformed mammalian cells, particularly myeloma cells of rodent, bovine and human origin. Usually, rat or mouse myeloma cell lines are employed. The hybridoma cells can be cultured in a suitable culture medium that preferably contains one or more substances that inhibit the growth or survival of the unfused, immortalized cells. For

example, if the parental cells lack the enzyme hypoxanthine guanine phosphoribosyl transferase (HGPRT or HPRT), the culture medium for the hybridomas typically will include hypoxanthine, aminopterin, and thymidine ("HAT medium"), which substances prevent the growth of HGPRT-deficient cells.

Preferred immortalized cell lines are those that fuse efficiently, support stable high level expression of antibody by the selected antibody-producing cells, and are sensitive to a medium such as HAT medium. More preferred immortalized cell lines are murine myeloma lines, which can be obtained, for instance, from the Salk Institute Cell Distribution Center, San Diego, California and the American Type Culture Collection, Manassas, Virginia. Human myeloma and mouse-human heteromyeloma cell lines also have been described for the production of human monoclonal antibodies (Kozbor, J. Immunol., 133:3001 (1984); Brodeur et al.. Monoclonal Antibody Production Techniques and Applications, Marcel Dekker, Inc., New York, (1987) pp. 51-63).

The culture medium in which the hybridoma cells are cultured can then be assayed for the presence of monoclonal antibodies directed against the antigen. Preferably, the binding specificity of monoclonal antibodies produced by the hybridoma cells is determined by immunoprecipitation or by an *in vitro* binding assay, such as radioimmunoassay (RIA) or enzyme-linked immunoabsorbent assay (ELISA). Such techniques and assays are known in the art. The binding affinity of the monoclonal antibody can, for example, be determined by the Scatchard analysis of Munson and Pollard, Anal. Biochem., 107:220 (1980). Preferably, antibodies having a high degree of specificity and a high binding affinity for the target antigen are isolated.

After the desired hybridoma cells are identified, the clones can be subcloned by limiting dilution procedures and grown by standard methods. Suitable culture media for this purpose include, for example, Dulbecco's Modified Eagle's Medium and RPMI-1640 medium. Alternatively, the hybridoma cells can be grown *in vivo* as ascites in a mammal. The monoclonal antibodies secreted by the subclones can be isolated or purified from the culture medium or ascites fluid by conventional immunoglobulin purification procedures such as, for example, protein A-Sepharose, hydroxylapatite chromatography, gel electrophoresis, dialysis, or affinity chromatography.

The monoclonal antibodies can also be made by recombinant DNA methods, such as those described in U.S. Patent No. 4,816,567. DNA encoding the monoclonal antibodies of the invention can be readily isolated and sequenced using conventional procedures (*e.g.*, by using oligonucleotide probes that are capable of binding specifically to genes encoding the heavy and light chains of murine antibodies). The hybridoma cells of the invention serve as a preferred source of such DNA. Once isolated, the DNA can be placed into expression vectors, which are then transfected into host cells such as simian COS cells, Chinese hamster ovary (CHO) cells, or myeloma cells that do not otherwise produce immunoglobulin protein, to obtain the synthesis of monoclonal antibodies in the recombinant host cells. The DNA also can be modified, for example, by substituting the coding sequence for human heavy and light chain constant domains in place of the homologous murine sequences (U.S. Patent No. 4,816,567; Morrison, Nature 368, 812-13 (1994)) or by covalently joining to the immunoglobulin coding sequence all or part of the coding sequence for a non-immunoglobulin polypeptide. Such a non-immunoglobulin polypeptide can be substituted for the constant domains of an antibody of the invention, or can be substituted for the variable domains of one antigen-combining site of an antibody of the invention to create a chimeric bivalent antibody.

5.13.2 HUMANIZED ANTIBODIES

The antibodies directed against the protein antigens of the invention can further comprise humanized antibodies or human antibodies. These antibodies are suitable for administration to humans without engendering an immune response by the human against the administered immunoglobulin. Humanized forms of antibodies are chimeric immunoglobulins, immunoglobulin chains or fragments thereof (such as Fv, Fab, Fab', $F(ab')_2$ or other antigen-binding subsequences of antibodies) that are principally comprised of the sequence of a human immunoglobulin, and contain minimal sequence derived from a non-human immunoglobulin. Humanization can be performed following the method of Winter and co-workers (Jones et al.. Nature, 321:522-525 (1986); Riechmann et al.. Nature, 332:323-327 (1988); Verhoeyen et al.. Science, 239:1534-1536 (1988)), by substituting rodent CDRs or CDR sequences for the corresponding sequences of a human antibody. (See also U.S. Patent No. 5,225,539.) In some instances, Fv

framework residues of the human immunoglobulin are replaced by corresponding non-human residues. Humanized antibodies can also comprise residues which are found neither in the recipient antibody nor in the imported CDR or framework sequences. In general, the humanized antibody will comprise substantially all of at least one, and typically two, variable domains, in which all or substantially all of the CDR regions correspond to those of a non-human immunoglobulin and all or substantially all of the framework regions are those of a human immunoglobulin consensus sequence. The humanized antibody optimally also will comprise at least a portion of an immunoglobulin constant region (Fc), typically that of a human immunoglobulin (Jones et al.. 1986; Riechmann et al.. 1988; and Presta, Curr. Op. Struct. Biol., 2:593-596 (1992)).

5.13.3 HUMAN ANTIBODIES

Fully human antibodies relate to antibody molecules in which essentially the entire sequences of both the light chain and the heavy chain, including the CDRs, arise from human genes. Such antibodies are termed "human antibodies", or "fully human antibodies" herein. Human monoclonal antibodies can be prepared by the trioma technique; the human B-cell hybridoma technique (see Kozbor, et al.. 1983 Immunol Today 4: 72) and the EBV hybridoma technique to produce human monoclonal antibodies (see Cole, et al.. 1985 In: MONOCLONAL ANTIBODIES AND CANCER THERAPY, Alan R. Liss, Inc., pp. 77-96). Human monoclonal antibodies may be utilized in the practice of the present invention and may be produced by using human hybridomas (see Cote, et al.. 1983. Proc Natl Acad Sci USA 80: 2026-2030) or by transforming human B-cells with Epstein Barr Virus *in vitro* (see Cole, et al.. 1985 In: MONOCLONAL ANTIBODIES AND CANCER THERAPY, Alan R. Liss, Inc., pp. 77-96).

In addition, human antibodies can also be produced using additional techniques, including phage display libraries (Hoogenboom and Winter, J. Mol. Biol., 227:381 (1991); Marks et al.. J. Mol. Biol., 222:581 (1991)). Similarly, human antibodies can be made by introducing human immunoglobulin loci into transgenic animals, *e.g.*, mice in which the endogenous immunoglobulin genes have been partially or completely inactivated. Upon challenge, human antibody production is observed, which closely resembles that seen in humans in all respects, including gene rearrangement, assembly,

and antibody repertoire. This approach is described, for example, in U.S. Patent Nos. 5,545,807; 5,545,806; 5,569,825; 5,625,126; 5,633,425; 5,661,016, and in Marks et. al., (Bio/Technology 10, 779-783 (1992)); Lonberg et. al., (Nature 368 856-859 (1994)); Morrison (Nature 368, 812-13 (1994)); Fishwild et al,(Nature Biotechnology 14, 845-51 (1996)); Neuberger (Nature Biotechnology 14, 826 (1996)); and Lonberg and Huszar (Intern. Rev. Immunol. 13 65-93 (1995)).

Human antibodies may additionally be produced using transgenic nonhuman animals which are modified so as to produce fully human antibodies rather than the animal's endogenous antibodies in response to challenge by an antigen. (See PCT publication WO94/02602). The endogenous genes encoding the heavy and light immunoglobulin chains in the nonhuman host have been incapacitated, and active loci encoding human heavy and light chain immunoglobulins are inserted into the host's genome. The human genes are incorporated, for example, using yeast artificial chromosomes containing the requisite human DNA segments. An animal which provides all the desired modifications is then obtained as progeny by crossbreeding intermediate transgenic animals containing fewer than the full complement of the modifications. The preferred embodiment of such a nonhuman animal is a mouse, and is termed the Xenomouse™ (Abgenix Inc., Freemont, CA) as disclosed in PCT publications WO 96/33735 and WO 96/34096. This animal produces B cells which secrete fully human immunoglobulins. The antibodies can be obtained directly from the animal after immunization with an immunogen of interest, as, for example, a preparation of a polyclonal antibody, or alternatively from immortalized B cells derived from the animal, such as hybridomas producing monoclonal antibodies. Additionally, the genes encoding the immunoglobulins with human variable regions can be recovered and expressed to obtain the antibodies directly, or can be further modified to obtain analogs of antibodies such as, for example, single chain Fv molecules.

An example of a method of producing a nonhuman host, exemplified as a mouse, lacking expression of an endogenous immunoglobulin heavy chain is disclosed in U.S. Patent No. 5,939,598. It can be obtained by a method including deleting the J segment genes from at least one endogenous heavy chain locus in an embryonic stem cell to prevent rearrangement of the locus and to prevent formation of a transcript of a

rearranged immunoglobulin heavy chain locus, the deletion being effected by a targeting vector containing a gene encoding a selectable marker; and producing from the embryonic stem cell a transgenic mouse whose somatic and germ cells contain the gene encoding the selectable marker.

A method for producing an antibody of interest, such as a human antibody, is disclosed in U.S. Patent No. 5,916,771. It includes introducing an expression vector that contains a nucleotide sequence encoding a heavy chain into one mammalian host cell in culture, introducing an expression vector containing a nucleotide sequence encoding a light chain into another mammalian host cell, and fusing the two cells to form a hybrid cell. The hybrid cell expresses an antibody containing the heavy chain and the light chain.

In a further improvement on this procedure, a method for identifying a clinically relevant epitope on an immunogen, and a correlative method for selecting an antibody that binds immunospecifically to the relevant epitope with high affinity, are disclosed in PCT publication WO 99/53049.

5.13.4 F_{ab} FRAGMENTS AND SINGLE CHAIN ANTIBODIES

According to the invention, techniques can be adapted for the production of single-chain antibodies specific to an antigenic protein of the invention (see *e.g.*, U.S. Patent No. 4,946,778). In addition, methods can be adapted for the construction of F_{ab} expression libraries (see *e.g.*, Huse, et al.. 1989 Science 246: 1275-1281) to allow rapid and effective identification of monoclonal F_{ab} fragments with the desired specificity for a protein or derivatives, fragments, analogs or homologs thereof. Antibody fragments that contain the idiotypes to a protein antigen may be produced by techniques known in the art including, but not limited to: (i) an $F(ab')_2$ fragment produced by pepsin digestion of an antibody molecule; (ii) an F_{ab} fragment generated by reducing the disulfide bridges of an $F_{(ab')2}$ fragment; (iii) an F_{ab} fragment generated by the treatment of the antibody molecule with papain and a reducing agent and (iv) F_v fragments.

5.13.5 BISPECIFIC ANTIBODIES

Bispecific antibodies are monoclonal, preferably human or humanized, antibodies that have binding specificities for at least two different antigens. In the present case, one of the binding specificities is for an antigenic protein of the invention. The second binding target is any other antigen, and advantageously is a cell-surface protein or receptor or receptor subunit.

Methods for making bispecific antibodies are known in the art. Traditionally, the recombinant production of bispecific antibodies is based on the co-expression of two immunoglobulin heavy-chain/light-chain pairs, where the two heavy chains have different specificities (Milstein and Cuello, Nature, 305:537-539 (1983)). Because of the random assortment of immunoglobulin heavy and light chains, these hybridomas (quadromas) produce a potential mixture of ten different antibody molecules, of which only one has the correct bispecific structure. The purification of the correct molecule is usually accomplished by affinity chromatography steps. Similar procedures are disclosed in WO 93/08829, published 13 May 1993, and in Traunecker *et al..*, 1991 *EMBO J.*, 10:3655-3659.

Antibody variable domains with the desired binding specificities (antibody-antigen combining sites) can be fused to immunoglobulin constant domain sequences. The fusion preferably is with an immunoglobulin heavy-chain constant domain, comprising at least part of the hinge, CH2, and CH3 regions. It is preferred to have the first heavy-chain constant region (CH1) containing the site necessary for light-chain binding present in at least one of the fusions. DNAs encoding the immunoglobulin heavy-chain fusions and, if desired, the immunoglobulin light chain, are inserted into separate expression vectors, and are co-transfected into a suitable host organism. For further details of generating bispecific antibodies see, for example, Suresh et al.. Methods in Enzymology, 121:210 (1986).

According to another approach described in WO 96/27011, the interface between a pair of antibody molecules can be engineered to maximize the percentage of heterodimers which are recovered from recombinant cell culture. The preferred interface comprises at least a part of the CH3 region of an antibody constant domain. In this method, one or more small amino acid side chains from the interface of the first antibody

molecule are replaced with larger side chains (e.g. tyrosine or tryptophan). Compensatory "cavities" of identical or similar size to the large side chain(s) are created on the interface of the second antibody molecule by replacing large amino acid side chains with smaller ones (e.g. alanine or threonine). This provides a mechanism for increasing the yield of the heterodimer over other unwanted end-products such as homodimers.

Bispecific antibodies can be prepared as full length antibodies or antibody fragments (e.g. $F(ab')_2$ bispecific antibodies). Techniques for generating bispecific antibodies from antibody fragments have been described in the literature. For example, bispecific antibodies can be prepared using chemical linkage. Brennan et al.. Science 229:81 (1985) describe a procedure wherein intact antibodies are proteolytically cleaved to generate $F(ab')_2$ fragments. These fragments are reduced in the presence of the dithiol complexing agent sodium arsenite to stabilize vicinal dithiols and prevent intermolecular disulfide formation. The Fab' fragments generated are then converted to thionitrobenzoate (TNB) derivatives. One of the Fab'-TNB derivatives is then reconverted to the Fab'-thiol by reduction with mercaptoethylamine and is mixed with an equimolar amount of the other Fab'-TNB derivative to form the bispecific antibody. The bispecific antibodies produced can be used as agents for the selective immobilization of enzymes.

Additionally, Fab' fragments can be directly recovered from E. coli and chemically coupled to form bispecific antibodies. Shalaby et al.. J. Exp. Med. 175:217-225 (1992) describe the production of a fully humanized bispecific antibody $F(ab')_2$ molecule. Each Fab' fragment was separately secreted from E. coli and subjected to directed chemical coupling *in vitro* to form the bispecific antibody. The bispecific antibody thus formed was able to bind to cells overexpressing the ErbB2 receptor and normal human T cells, as well as trigger the lytic activity of human cytotoxic lymphocytes against human breast tumor targets.

Various techniques for making and isolating bispecific antibody fragments directly from recombinant cell culture have also been described. For example, bispecific antibodies have been produced using leucine zippers. Kostelny et al.. J. Immunol. 148(5):1547-1553 (1992). The leucine zipper peptides from the Fos and Jun proteins

were linked to the Fab' portions of two different antibodies by gene fusion. The antibody homodimers were reduced at the hinge region to form monomers and then re-oxidized to form the antibody heterodimers. This method can also be utilized for the production of antibody homodimers. The "diabody" technology described by Hollinger et al.. Proc. Natl. Acad. Sci. USA 90:6444-6448 (1993) has provided an alternative mechanism for making bispecific antibody fragments. The fragments comprise a heavy-chain variable domain (V_H) connected to a light-chain variable domain (V_L) by a linker which is too short to allow pairing between the two domains on the same chain. Accordingly, the V_H and V_L domains of one fragment are forced to pair with the complementary V_L and V_H domains of another fragment, thereby forming two antigen-binding sites. Another strategy for making bispecific antibody fragments by the use of single-chain Fv (sFv) dimers has also been reported. See, Gruber et al.. J. Immunol. 152:5368 (1994).

Antibodies with more than two valencies are contemplated. For example, trispecific antibodies can be prepared. Tutt et al.. J. Immunol. 147:60 (1991). Exemplary bispecific antibodies can bind to two different epitopes, at least one of which originates in the protein antigen of the invention. Alternatively, an anti-antigenic arm of an immunoglobulin molecule can be combined with an arm which binds to a triggering molecule on a leukocyte such as a T-cell receptor molecule (e.g. CD2, CD3, CD28, or B7), or Fc receptors for IgG (FcγR), such as FcRn, FcγRI (CD64), FcγRII (CD32), FcγRIII (CD16), so as to focus cellular defense mechanisms to the cell expressing the particular antigen. Bispecific antibodies can also be used to direct cytotoxic agents to cells which express a particular antigen. These antibodies possess an antigen-binding arm and an arm which binds a cytotoxic agent or a radionuclide chelator, such as EOTUBE, DPTA, DOTA, or TETA. Another bispecific antibody of interest binds the protein antigen described herein and further binds tissue factor (TF).

5.13.6 HETEROCONJUGATE ANTIBODIES

Heteroconjugate antibodies are also within the scope of the present invention. Heteroconjugate antibodies are composed of two covalently joined antibodies. Such antibodies have, for example, been proposed to target immune system cells to unwanted cells (U.S. Patent No. 4,676,980), and for treatment of HIV infection (WO 91/00360;

WO 92/200373; EP 03089). It is contemplated that the antibodies can be prepared *in vitro* using known methods in synthetic protein chemistry, including those involving crosslinking agents. For example, immunotoxins can be constructed using a disulfide exchange reaction or by forming a thioether bond. Examples of suitable reagents for this purpose include iminothiolate and methyl-4-mercaptobutyrimidate and those disclosed, for example, in U.S. Patent No. 4,676,980.

5.13.7 EFFECTOR FUNCTION ENGINEERING

It can be desirable to modify the antibody of the invention with respect to effector function, so as to enhance, *e.g.,* the effectiveness of the antibody in treating cancer. For example, cysteine residue(s) can be introduced into the Fc region, thereby allowing interchain disulfide bond formation in this region. The homodimeric antibody thus generated can have improved internalization capability and/or increased complement-mediated cell killing and antibody-dependent cellular cytotoxicity (ADCC). See Caron et al.. J. Exp Med., 176: 1191-1195 (1992) and Shopes, J. Immunol., 148: 2918-2922 (1992). Homodimeric antibodies with enhanced anti-tumor activity can also be prepared using heterobifunctional cross-linkers as described in Wolff et. al., Cancer Research, 53: 2560-2565 (1993). Alternatively, an antibody can be engineered that has dual Fc regions and can thereby have enhanced complement lysis and ADCC capabilities. See Stevenson et al.. Anti-Cancer Drug Design, 3: 219-230 (1989).

5.13.8 IMMUNOCONJUGATES

The invention also pertains to immunoconjugates comprising an antibody conjugated to a cytotoxic agent such as a chemotherapeutic agent, toxin (*e.g.,* an enzymatically active toxin of bacterial, fungal, plant, or animal origin, or fragments thereof), or a radioactive isotope (*i.e.,* a radioconjugate).

Chemotherapeutic agents useful in the generation of such immunoconjugates have been described above. Enzymatically active toxins and fragments thereof that can be used include diphtheria A chain, nonbinding active fragments of diphtheria toxin, exotoxin A chain (from Pseudomonas aeruginosa), ricin A chain, abrin A chain, modeccin A chain, alpha-sarcin, Aleurites fordii proteins, dianthin proteins, Phytolaca

americana proteins (PAPI, PAPII, and PAP-S), momordica charantia inhibitor, curcin, crotin, sapaonaria officinalis inhibitor, gelonin, mitogellin, restrictocin, phenomycin, enomycin, and the tricothecenes. A variety of radionuclides are available for the production of radioconjugated antibodies. Examples include ^{212}Bi, ^{131}I, ^{125}I, ^{131}In, ^{90}Y, and ^{186}Re.

Conjugates of the antibody and cytotoxic agent are made using a variety of bifunctional protein-coupling agents such as N-succinimidyl-3-(2-pyridyldithiol) propionate (SPDP), iminothiolane (IT), bifunctional derivatives of imidoesters (such as dimethyl adipimidate HCL), active esters (such as disuccinimidyl suberate), aldehydes (such as glutareldehyde), bis-azido compounds (such as bis (p-azidobenzoyl) hexanediamine), bis-diazonium derivatives (such as bis-(p-diazoniumbenzoyl)-ethylenediamine), diisocyanates (such as tolyene 2,6-diisocyanate), and bis-active fluorine compounds (such as 1,5-difluoro-2,4-dinitrobenzene). For example, a ricin immunotoxin can be prepared as described in Vitetta et al.. Science, 238: 1098 (1987). Carbon-14-labeled 1-isothiocyanatobenzyl-3-methyldiethylene triaminepentaacetic acid (MX-DTPA) is an exemplary chelating agent for conjugation of radionucleotide to the antibody. See WO94/11026.

In another embodiment, the antibody can be conjugated to a "receptor" (such streptavidin) for utilization in tumor pretargeting wherein the antibody-receptor conjugate is administered to the patient, followed by removal of unbound conjugate from the circulation using a clearing agent and then administration of a "ligand" (*e.g.*, avidin) that is in turn conjugated to a cytotoxic agent.

5.14 COMPUTER READABLE SEQUENCES

In one application of this embodiment, a nucleotide sequence of the present invention can be recorded on computer readable media. As used herein, "computer readable media" refers to any medium which can be read and accessed directly by a computer. Such media include, but are not limited to: magnetic storage media, such as floppy discs, hard disc storage medium, and magnetic tape; optical storage media such as CD-ROM; electrical storage media such as RAM and ROM; and hybrids of these categories such as magnetic/optical storage media. A skilled artisan can readily

appreciate how any of the presently known computer readable mediums can be used to create a manufacture comprising computer readable medium having recorded thereon a nucleotide sequence of the present invention. As used herein, "recorded" refers to a process for storing information on computer readable medium. A skilled artisan can readily adopt any of the presently known methods for recording information on computer readable medium to generate manufactures comprising the nucleotide sequence information of the present invention.

A variety of data storage structures are available to a skilled artisan for creating a computer readable medium having recorded thereon a nucleotide sequence of the present invention. The choice of the data storage structure will generally be based on the means chosen to access the stored information. In addition, a variety of data processor programs and formats can be used to store the nucleotide sequence information of the present invention on computer readable medium. The sequence information can be represented in a word processing text file, formatted in commercially-available software such as WordPerfect and Microsoft Word, or represented in the form of an ASCII file, stored in a database application, such as DB2, Sybase, Oracle, or the like. A skilled artisan can readily adapt any number of data processor structuring formats (*e.g.* text file or database) in order to obtain computer readable medium having recorded thereon the nucleotide sequence information of the present invention.

By providing any of the nucleotide sequences 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502 or a representative fragment thereof; or a nucleotide sequence at least 95% identical to any of the nucleotide sequences of 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502 in computer readable form, a skilled artisan can routinely access the sequence information for a variety of purposes. Computer software is publicly available

which allows a skilled artisan to access sequence information provided in a computer readable medium. The examples which follow demonstrate how software which implements the BLAST (Altschul et al.. J. Mol. Biol. 215:403-410 (1990)) and BLAZE (Brutlag et al.. Comp. Chem. 17:203-207 (1993)) search algorithms on a Sybase system is used to identify open reading frames (ORFs) within a nucleic acid sequence. Such ORFs may be protein encoding fragments and may be useful in producing commercially important proteins such as enzymes used in fermentation reactions and in the production of commercially useful metabolites.

As used herein, "a computer-based system" refers to the hardware means, software means, and data storage means used to analyze the nucleotide sequence information of the present invention. The minimum hardware means of the computer-based systems of the present invention comprises a central processing unit (CPU), input means, output means, and data storage means. A skilled artisan can readily appreciate that any one of the currently available computer-based systems are suitable for use in the present invention. As stated above, the computer-based systems of the present invention comprise a data storage means having stored therein a nucleotide sequence of the present invention and the necessary hardware means and software means for supporting and implementing a search means. As used herein, "data storage means" refers to memory which can store nucleotide sequence information of the present invention, or a memory access means which can access manufactures having recorded thereon the nucleotide sequence information of the present invention.

As used herein, "search means" refers to one or more programs which are implemented on the computer-based system to compare a target sequence or target structural motif with the sequence information stored within the data storage means. Search means are used to identify fragments or regions of a known sequence which match a particular target sequence or target motif. A variety of known algorithms are disclosed publicly and a variety of commercially available software for conducting search means are and can be used in the computer-based systems of the present invention. Examples of such software includes, but is not limited to, Smith-Waterman, MacPattern (EMBL), BLASTN and BLASTA (NPOLYPEPTIDEIA). A skilled artisan can readily recognize that any one of the available algorithms or implementing software packages for

conducting homology searches can be adapted for use in the present computer-based systems. As used herein, a "target sequence" can be any nucleic acid or amino acid sequence of six or more nucleotides or two or more amino acids. A skilled artisan can readily recognize that the longer a target sequence is, the less likely a target sequence will be present as a random occurrence in the database. The most preferred sequence length of a target sequence is from about 10 to 300 amino acids, more preferably from about 30 to 100 nucleotide residues. However, it is well recognized that searches for commercially important fragments, such as sequence fragments involved in gene expression and protein processing, may be of shorter length.

As used herein, "a target structural motif," or "target motif," refers to any rationally selected sequence or combination of sequences in which the sequence(s) are chosen based on a three-dimensional configuration which is formed upon the folding of the target motif. There are a variety of target motifs known in the art. Protein target motifs include, but are not limited to, enzyme active sites and signal sequences. Nucleic acid target motifs include, but are not limited to, promoter sequences, hairpin structures and inducible expression elements (protein binding sequences).

5.15 TRIPLE HELIX FORMATION

In addition, the fragments of the present invention, as broadly described, can be used to control gene expression through triple helix formation or antisense DNA or RNA, both of which methods are based on the binding of a polynucleotide sequence to DNA or RNA. Polynucleotides suitable for use in these methods are preferably 20 to 40 bases in length and are designed to be complementary to a region of the gene involved in transcription (triple helix - see Lee et al.. Nucl. Acids Res. 6:3073 (1979); Cooney et al.. Science 15241:456 (1988); and Dervan et al.. Science 251:1360 (1991)) or to the mRNA itself (antisense - Olmno, J. Neurochem. 56:560 (1991); Oligodeoxynucleotides as Antisense Inhibitors of Gene Expression, CRC Press, Boca Raton, FL (1988)). Triple helix-formation optimally results in a shut-off of RNA transcription from DNA, while antisense RNA hybridization blocks translation of an mRNA molecule into polypeptide. Both techniques have been demonstrated to be effective in model systems. Information

contained in the sequences of the present invention is necessary for the design of an antisense or triple helix oligonucleotide.

5.16 DIAGNOSTIC ASSAYS AND KITS

The present invention further provides methods to identify the presence or expression of one of the ORFs of the present invention, or homolog thereof, in a test sample, using a nucleic acid probe or antibodies of the present invention, optionally conjugated or otherwise associated with a suitable label.

In general, methods for detecting a polynucleotide of the invention can comprise contacting a sample with a compound that binds to and forms a complex with the polynucleotide for a period sufficient to form the complex, and detecting the complex, so that if a complex is detected, a polynucleotide of the invention is detected in the sample. Such methods can also comprise contacting a sample under stringent hybridization conditions with nucleic acid primers that anneal to a polynucleotide of the invention under such conditions, and amplifying annealed polynucleotides, so that if a polynucleotide is amplified, a polynucleotide of the invention is detected in the sample.

In general, methods for detecting a polypeptide of the invention can comprise contacting a sample with a compound that binds to and forms a complex with the polypeptide for a period sufficient to form the complex, and detecting the complex, so that if a complex is detected, a polypeptide of the invention is detected in the sample.

In detail, such methods comprise incubating a test sample with one or more of the antibodies or one or more of the nucleic acid probes of the present invention and assaying for binding of the nucleic acid probes or antibodies to components within the test sample.

Conditions for incubating a nucleic acid probe or antibody with a test sample vary. Incubation conditions depend on the format employed in the assay, the detection methods employed, and the type and nature of the nucleic acid probe or antibody used in the assay. One skilled in the art will recognize that any one of the commonly available hybridization, amplification or immunological assay formats can readily be adapted to employ the nucleic acid probes or antibodies of the present invention. Examples of such assays can be found in Chard, T., An Introduction to Radioimmunoassay and Related Techniques, Elsevier Science Publishers, Amsterdam, The Netherlands (1986); Bullock,

G.R. et al.. Techniques in Immunocytochemistry, Academic Press, Orlando, FL Vol. 1 (1982), Vol. 2 (1983), Vol. 3 (1985); Tijssen, P., Practice and Theory of immunoassays: Laboratory Techniques in Biochemistry and Molecular Biology, Elsevier Science Publishers, Amsterdam, The Netherlands (1985). The test samples of the present invention include cells, protein or membrane extracts of cells, or biological fluids such as sputum, blood, serum, plasma, or urine. The test sample used in the above-described method will vary based on the assay format, nature of the detection method and the tissues, cells or extracts used as the sample to be assayed. Methods for preparing protein extracts or membrane extracts of cells are well known in the art and can be readily be adapted in order to obtain a sample which is compatible with the system utilized.

In another embodiment of the present invention, kits are provided which contain the necessary reagents to carry out the assays of the present invention. Specifically, the invention provides a compartment kit to receive, in close confinement, one or more containers which comprises: (a) a first container comprising one of the probes or antibodies of the present invention; and (b) one or more other containers comprising one or more of the following: wash reagents, reagents capable of detecting presence of a bound probe or antibody.

In detail, a compartment kit includes any kit in which reagents are contained in separate containers. Such containers include small glass containers, plastic containers or strips of plastic or paper. Such containers allows one to efficiently transfer reagents from one compartment to another compartment such that the samples and reagents are not cross-contaminated, and the agents or solutions of each container can be added in a quantitative fashion from one compartment to another. Such containers will include a container which will accept the test sample, a container which contains the antibodies used in the assay, containers which contain wash reagents (such as phosphate buffered saline, Tris-buffers, etc.), and containers which contain the reagents used to detect the bound antibody or probe. Types of detection reagents include labeled nucleic acid probes, labeled secondary antibodies, or in the alternative, if the primary antibody is labeled, the enzymatic, or antibody binding reagents which are capable of reacting with the labeled antibody. One skilled in the art will readily recognize that the disclosed probes and

antibodies of the present invention can be readily incorporated into one of the established kit formats which are well known in the art.

5.17 MEDICAL IMAGING

The novel polypeptides and binding partners of the invention are useful in medical imaging of sites expressing the molecules of the invention (*e.g.*, where the polypeptide of the invention is involved in the immune response, for imaging sites of inflammation or infection). See, *e.g.*, Kunkel et al.. U.S. Pat. NO. 5,413,778. Such methods involve chemical attachment of a labeling or imaging agent, administration of the labeled polypeptide to a subject in a pharmaceutically acceptable carrier, and imaging the labeled polypeptide *in vivo* at the target site.

5.18 SCREENING ASSAYS

Using the isolated proteins and polynucleotides disclosed herein, the present invention further provides methods of obtaining and identifying agents which bind to a polypeptide encoded by an ORF corresponding to any of the nucleotide sequences set forth in 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502, or bind to a specific domain of the polypeptide encoded by the nucleic acid. In detail, said method comprises the steps of:

(a) contacting an agent with an isolated protein encoded by an ORF of the present invention, or nucleic acid of the invention; and

(b) determining whether the agent binds to said protein or said nucleic acid.

In general, therefore, such methods for identifying compounds that bind to a polynucleotide of the invention can comprise contacting a compound with a polynucleotide of the invention for a time sufficient to form a polynucleotide/compound complex, and detecting the complex, so that if a polynucleotide/compound complex is detected, a compound that binds to a polynucleotide of the invention is identified.

Likewise, in general, therefore, such methods for identifying compounds that bind to a polypeptide of the invention can comprise contacting a compound with a polypeptide of the invention for a time sufficient to form a polypeptide/compound complex, and detecting the complex, so that if a polypeptide/compound complex is detected, a compound that binds to a polynucleotide of the invention is identified.

Methods for identifying compounds that bind to a polypeptide of the invention can also comprise contacting a compound with a polypeptide of the invention in a cell for a time sufficient to form a polypeptide/compound complex, wherein the complex drives expression of a receptor gene sequence in the cell, and detecting the complex by detecting reporter gene sequence expression, so that if a polypeptide/compound complex is detected, a compound that binds a polypeptide of the invention is identified.

Compounds identified via such methods can include compounds which modulate the activity of a polypeptide of the invention (that is, increase or decrease its activity, relative to activity observed in the absence of the compound). Alternatively, compounds identified via such methods can include compounds which modulate the expression of a polynucleotide of the invention (that is, increase or decrease expression relative to expression levels observed in the absence of the compound). Compounds, such as compounds identified via the methods of the invention, can be tested using standard assays well known to those of skill in the art for their ability to modulate activity/expression.

The agents screened in the above assay can be, but are not limited to, peptides, carbohydrates, vitamin derivatives, or other pharmaceutical agents. The agents can be selected and screened at random or rationally selected or designed using protein modeling techniques.

For random screening, agents such as peptides, carbohydrates, pharmaceutical agents and the like are selected at random and are assayed for their ability to bind to the protein encoded by the ORF of the present invention. Alternatively, agents may be rationally selected or designed. As used herein, an agent is said to be "rationally selected or designed" when the agent is chosen based on the configuration of the particular protein. For example, one skilled in the art can readily adapt currently available procedures to generate peptides, pharmaceutical agents and the like, capable of binding to

a specific peptide sequence, in order to generate rationally designed antipeptide peptides, for example see Hurby et al.. Application of Synthetic Peptides: Antisense Peptides," In Synthetic Peptides, A User's Guide, W.H. Freeman, NY (1992), pp. 289-307, and Kaspczak et al.. Biochemistry 28:9230-8 (1989), or pharmaceutical agents, or the like.

In addition to the foregoing, one class of agents of the present invention, as broadly described, can be used to control gene expression through binding to one of the ORFs or EMFs of the present invention. As described above, such agents can be randomly screened or rationally designed/selected. Targeting the ORF or EMF allows a skilled artisan to design sequence specific or element specific agents, modulating the expression of either a single ORF or multiple ORFs which rely on the same EMF for expression control. One class of DNA binding agents are agents which contain base residues which hybridize or form a triple helix formation by binding to DNA or RNA. Such agents can be based on the classic phosphodiester, ribonucleic acid backbone, or can be a variety of sulfhydryl or polymeric derivatives which have base attachment capacity.

Agents suitable for use in these methods preferably contain 20 to 40 bases and are designed to be complementary to a region of the gene involved in transcription (triple helix - see Lee et al.. Nucl. Acids Res. 6:3073 (1979); Cooney et al.. Science 241:456 (1988); and Dervan et al.. Science 251:1360 (1991)) or to the mRNA itself (antisense - Okano, J. Neurochem. 56:560 (1991); Oligodeoxynucleotides as Antisense Inhibitors of Gene Expression, CRC Press, Boca Raton, FL (1988)). Triple helix-formation optimally results in a shut-off of RNA transcription from DNA, while antisense RNA hybridization blocks translation of an mRNA molecule into polypeptide. Both techniques have been demonstrated to be effective in model systems. Information contained in the sequences of the present invention is necessary for the design of an antisense or triple helix oligonucleotide and other DNA binding agents.

Agents which bind to a protein encoded by one of the ORFs of the present invention can be used as a diagnostic agent. Agents which bind to a protein encoded by one of the ORFs of the present invention can be formulated using known techniques to generate a pharmaceutical composition.

5.19 USE OF NUCLEIC ACIDS AS PROBES

Another aspect of the present invention is to provide for polypeptide-specific nucleic acid hybridization probes capable of hybridizing with naturally occurring nucleotide sequences. The hybridization probes of the subject invention may be derived from any of the nucleotide sequences 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502. Because the corresponding gene is only expressed in a limited number of tissues, a hybridization probe derived from of any of the nucleotide sequences 748 SEQ ID NO: 1-45,207, 752 SEQ ID NO: 1-13,203, 778 SEQ ID NO: 1-105, 779 SEQ ID NO: 1-128, 782 SEQ ID NO: 1-10,451, 784 SEQ ID NO: 1-10,289, 785 SEQ ID NO: 1-3796, 787 SEQ ID NO: 1-10,410, 788 SEQ ID NO: 1-14,074, 789 SEQ ID NO: 1-6391, 790 SEQ ID NO: 1-30,533, 791 SEQ ID NO: 1-5822, and 792 SEQ ID NO: 1-8502 can be used as an indicator of the presence of RNA of cell type of such a tissue in a sample.

Any suitable hybridization technique can be employed, such as, for example, *in situ* hybridization. PCR as described in US Patents Nos. 4,683,195 and 4,965,188 provides additional uses for oligonucleotides based upon the nucleotide sequences. Such probes used in PCR may be of recombinant origin, may be chemically synthesized, or a mixture of both. The probe will comprise a discrete nucleotide sequence for the detection of identical sequences or a degenerate pool of possible sequences for identification of closely related genomic sequences.

Other means for producing specific hybridization probes for nucleic acids include the cloning of nucleic acid sequences into vectors for the production of mRNA probes. Such vectors are known in the art and are commercially available and may be used to synthesize RNA probes *in vitro* by means of the addition of the appropriate RNA polymerase as T7 or SP6 RNA polymerase and the appropriate radioactively labeled nucleotides. The nucleotide sequences may be used to construct hybridization probes for mapping their respective genomic sequences. The nucleotide sequence provided herein may be mapped to a chromosome or specific regions of a chromosome using well known

genetic and/or chromosomal mapping techniques. These techniques include *in situ* hybridization, linkage analysis against known chromosomal markers, hybridization screening with libraries or flow-sorted chromosomal preparations specific to known chromosomes, and the like. The technique of fluorescent *in situ* hybridization of chromosome spreads has been described, among other places, in Verma et al (1988) Human Chromosomes: A Manual of Basic Techniques, Pergamon Press, New York NY.

Fluorescent *in situ* hybridization of chromosomal preparations and other physical chromosome mapping techniques may be correlated with additional genetic map data. Examples of genetic map data can be found in the 1994 Genome Issue of Science (265:1981f). Correlation between the location of a nucleic acid on a physical chromosomal map and a specific disease (or predisposition to a specific disease) may help delimit the region of DNA associated with that genetic disease. The nucleotide sequences of the subject invention may be used to detect differences in gene sequences between normal, carrier or affected individuals.

5.20 PREPARATION OF SUPPORT BOUND OLIGONUCLEOTIDES

Oligonucleotides, *i.e.,* small nucleic acid segments, may be readily prepared from the sequences disclosed herein by, for example, directly synthesizing the oligonucleotide by chemical means, as is commonly practiced using an automated oligonucleotide synthesizer.

Support bound oligonucleotides may be prepared by any of the methods known to those of skill in the art using any suitable support such as glass, polystyrene or Teflon® (DuPont). One strategy is to precisely spot oligonucleotides synthesized by standard synthesizers. Immobilization can be achieved using passive adsorption (Inouye & Hondo, (1990) J. Clin. Microbiol. 28(6) 1469-72); using UV light (Nagata *et al.*. 1985; Dahlen *et al.*. 1987; Morrissey & Collins, (1989) Mol. Cell Probes 3(2) 189-207) or by covalent binding of base modified DNA (Keller *et al.*. 1988; 1989); all references being specifically incorporated herein.

Another strategy that may be employed is the use of the strong biotin-streptavidin interaction as a linker. For example, Broude *et al.*. (1994) Proc. Natl. Acad. Sci. USA 91(8) 3072-6, describe the use of biotinylated probes, although these are duplex probes, that are immobilized on streptavidin-coated magnetic beads. Streptavidin-coated beads may be purchased from Dynal, Oslo. Of course, this same linking chemistry is applicable to coating

any surface with streptavidin. Biotinylated probes may be purchased from various sources, such as, *e.g.,* Operon Technologies (Alameda, CA).

Nunc Laboratories (Naperville, IL) is also selling suitable material that could be used. Nunc Laboratories have developed a method by which DNA can be covalently bound to the microwell surface termed Covalink NH. CovaLink NH is a polystyrene surface grafted with secondary amino groups (>NH) that serve as bridge-heads for further covalent coupling. CovaLink Modules may be purchased from Nunc Laboratories. DNA molecules may be bound to CovaLink exclusively at the 5'-end by a phosphoramidate bond, allowing immobilization of more than 1 pmol of DNA (Rasmussen *et al.*. (1991) Anal. Biochem. 198(1) 138-42).

The use of CovaLink NH strips for covalent binding of DNA molecules at the 5'-end has been described (Rasmussen et al.. (1991). In this technology, a phosphoramidate bond is employed (Chu et al.. (1983) Nucleic Acids Res. 11(8) 6513-29). This is beneficial as immobilization using only a single covalent bond is preferred. The phosphoramidate bond joins the DNA to the CovaLink NH secondary amino groups that are positioned at the end of spacer arms covalently grafted onto the polystyrene surface through a 2 nm long spacer arm. To link an oligonucleotide to CovaLink NH via an phosphoramidate bond, the oligonucleotide terminus must have a 5'-end phosphate group. It is, perhaps, even possible for biotin to be covalently bound to CovaLink and then streptavidin used to bind the probes.

More specifically, the linkage method includes dissolving DNA in water (7.5 ng/ul) and denaturing for 10 min. at 95°C and cooling on ice for 10 min. Ice-cold 0.1 M 1-methylimidazole, pH 7.0 (1-$MeIm_7$), is then added to a final concentration of 10 mM 1-$MeIm_7$. A ss DNA solution is then dispensed into CovaLink NH strips (75 ul/well) standing on ice.

Carbodiimide 0.2 M 1-ethyl-3-(3-dimethylaminopropyl)-carbodiimide (EDC), dissolved in 10 mM 1-$MeIm_7$, is made fresh and 25 ul added per well. The strips are incubated for 5 hours at 50°C. After incubation the strips are washed using, *e.g.,* Nunc-Immuno Wash; first the wells are washed 3 times, then they are soaked with washing solution for 5 min., and finally they are washed 3 times (where in the washing solution is 0.4 N NaOH, 0.25% SDS heated to 50°C).

It is contemplated that a further suitable method for use with the present invention is that described in PCT Patent Application WO 90/03382 (Southern & Maskos), incorporated herein by reference. This method of preparing an oligonucleotide bound to a support involves attaching a nucleoside 3'-reagent through the phosphate group by a covalent phosphodiester link to aliphatic hydroxyl groups carried by the support. The oligonucleotide is then synthesized on the supported nucleoside and protecting groups removed from the synthetic oligonucleotide chain under standard conditions that do not cleave the oligonucleotide from the support. Suitable reagents include nucleoside phosphoramidite and nucleoside hydrogen phosphorate.

An on-chip strategy for the preparation of DNA probe for the preparation of DNA probe arrays may be employed. For example, addressable laser-activated photodeprotection may be employed in the chemical synthesis of oligonucleotides directly on a glass surface, as described by Fodor *et al.*. (1991) Science 251(4995) 767-73, incorporated herein by reference. Probes may also be immobilized on nylon supports as described by Van Ness *et al.*. (1991) Nucleic Acids Res. 19(12) 3345-50; or linked to Teflon using the method of Duncan & Cavalier (1988) Anal. Biochem. 169(1) 104-8; all references being specifically incorporated herein.

To link an oligonucleotide to a nylon support, as described by Van Ness *et al.*. (1991), requires activation of the nylon surface via alkylation and selective activation of the 5'-amine of oligonucleotides with cyanuric chloride.

One particular way to prepare support bound oligonucleotides is to utilize the light-generated synthesis described by Pease *et al.*. (1994) PNAS USA 91(11) 5022-6, incorporated herein by reference). These authors used current photolithographic techniques to generate arrays of immobilized oligonucleotide probes (DNA chips). These methods, in which light is used to direct the synthesis of oligonucleotide probes in high-density, miniaturized arrays, utilize photolabile 5'-protected *N*-acyl-deoxynucleoside phosphoramidites, surface linker chemistry and versatile combinatorial synthesis strategies. A matrix of 256 spatially defined oligonucleotide probes may be generated in this manner.

5.21 PREPARATION OF NUCLEIC ACID FRAGMENTS

The nucleic acids may be obtained from any appropriate source, such as cDNAs, genomic DNA, chromosomal DNA, microdissected chromosome bands, cosmid or YAC

inserts, and RNA, including mRNA without any amplification steps. For example, Sambrook *et al.*. (1989) describes three protocols for the isolation of high molecular weight DNA from mammalian cells (p. 9.14-9.23).

DNA fragments may be prepared as clones in M13, plasmid or lambda vectors and/or prepared directly from genomic DNA or cDNA by PCR or other amplification methods. Samples may be prepared or dispensed in multiwell plates. About 100-1000 ng of DNA samples may be prepared in 2-500 ml of final volume.

The nucleic acids would then be fragmented by any of the methods known to those of skill in the art including, for example, using restriction enzymes as described at 9.24-9.28 of Sambrook *et al.*. (1989), shearing by ultrasound and NaOH treatment.

Low pressure shearing is also appropriate, as described by Schriefer *et al.*. (1990) Nucleic Acids Res. 18(24) 7455-6, incorporated herein by reference). In this method, DNA samples are passed through a small French pressure cell at a variety of low to intermediate pressures. A lever device allows controlled application of low to intermediate pressures to the cell. The results of these studies indicate that low-pressure shearing is a useful alternative to sonic and enzymatic DNA fragmentation methods.

One particularly suitable way for fragmenting DNA is contemplated to be that using the two base recognition endonuclease, *Cvi*JI, described by Fitzgerald *et al.*. (1992) Nucleic Acids Res. 20(14) 3753-62. These authors described an approach for the rapid fragmentation and fractionation of DNA into particular sizes that they contemplated to be suitable for shotgun cloning and sequencing.

The restriction endonuclease *Cvi*JI normally cleaves the recognition sequence PuGCPy between the G and C to leave blunt ends. Atypical reaction conditions, which alter the specificity of this enzyme (*Cvi*JI**), yield a quasi-random distribution of DNA fragments form the small molecule pUC19 (2688 base pairs). Fitzgerald *et al.*. (1992) quantitatively evaluated the randomness of this fragmentation strategy, using a *Cvi*JI** digest of pUC19 that was size fractionated by a rapid gel filtration method and directly ligated, without end repair, to a lac Z minus M13 cloning vector. Sequence analysis of 76 clones showed that *Cvi*JI** restricts pyGCPy and PuGCPu, in addition to PuGCPy sites, and that new sequence data is accumulated at a rate consistent with random fragmentation.

As reported in the literature, advantages of this approach compared to sonication and agarose gel fractionation include: smaller amounts of DNA are required (0.2-0.5 μg instead of 2-5 μg); and fewer steps are involved (no preligation, end repair, chemical extraction, or agarose gel electrophoresis and elution are needed

Irrespective of the manner in which the nucleic acid fragments are obtained or prepared, it is important to denature the DNA to give single stranded pieces available for hybridization. This is achieved by incubating the DNA solution for 2-5 minutes at 80-90°C. The solution is then cooled quickly to 2°C to prevent renaturation of the DNA fragments before they are contacted with the chip. Phosphate groups must also be removed from genomic DNA by methods known in the art.

5.22 PREPARATION OF DNA ARRAYS

The nucleic acid sequences disclose herein can be used to create arrays. Arrays may be prepared by spotting DNA samples on a support such as a nylon membrane. Spotting may be performed by using arrays of metal pins (the positions of which correspond to an array of wells in a microtiter plate) to repeated by transfer of about 20 nl of a DNA solution to a nylon membrane. By offset printing, a density of dots higher than the density of the wells is achieved. One to 25 dots may be accommodated in 1 mm^2, depending on the type of label used. By avoiding spotting in some preselected number of rows and columns, separate subsets (subarrays) may be formed. Samples in one subarray may be the same genomic segment of DNA (or the same gene) from different individuals, or may be different, overlapped genomic clones. Each of the subarrays may represent replica spotting of the same samples. In one example, a selected gene segment may be amplified from 64 patients. For each patient, the amplified gene segment may be in one 96-well plate (all 96 wells containing the same sample). A plate for each of the 64 patients is prepared. By using a 96-pin device, all samples may be spotted on one 8 x 12 cm membrane. Subarrays may contain 64 samples, one from each patient. Where the 96 subarrays are identical, the dot span may be 1 mm^2 and there may be a 1 mm space between subarrays.

Another approach is to use membranes or plates (available from NUNC, Naperville, Illinois) which may be partitioned by physical spacers e.g. a plastic grid molded over the membrane, the grid being similar to the sort of membrane applied to the bottom of multiwell

plates, or hydrophobic strips. A fixed physical spacer is not preferred for imaging by exposure to flat phosphor-storage screens or x-ray films.

5.23 PREPARATION OF A UNIVERSAL SET OF PROBES

The nucleic acid sequences disclosed herein may be used to design universal hybidizaiton probes for detecting genetic markers. Two types of universal sets of probes may be prepared. The first is a complete set (or at least a noncomplementary subset) of relatively short probes, for example all 4096 (or about 2000 non-complementary) 6-mers, or all 16,384 (or about 8,000 non-complementary) 7-mers. Full noncomplementary subsets of 8-mers and longer probes are less convenient inasmuch as they include 32,000 or more probes.

A second type of probe set is selected as a small subset of probes still sufficient for reading every bp in any sequence with at least with one probe. For example, 12 of 16 dimers are sufficient. A small subset for 7-mers, 8-mer and 9-mers for sequencing double stranded DNA may be about 3000, 10,000 and 30,000 probes, respectively.

Probes may be prepared using standard chemistry with one to three non-specified (mixed A,T,C and El) or universal (e.g. M base or inosine) bases at the ends. If radiolabelling is used, probes may have an OH group at the 5' end for kinasing by radiolabelled phosphorous groups. Alternatively, probes labelled with any compatible system, such as fluorescent dyes, may be employed. Other types of probes, such as PNA (Protein Nucleic Acids) or probes containing modified bases which change duplex stability also may be used.

Probes may be stored in bar-coded multiwell plates. For small numbers of probes, 96-well plates may be used; for 10,000 or more probes, storage in 384- or 864-well plates is preferred. Stacks of 5 to 50 plates are enough to store all probes. Approximately 5 pg of a probe may be sufficient for hybridization with one DNA sample. Thus, from a small synthesis of about 50 mg per probe, ten million samples may be analyzed. If each probe is used for every third sample, and if each sample is 1000 bp in length, then over 30 billion bases (10 human genomes) may be sequenced by a set of 5,000 probes.

5.24 PROBES HAVING MODIFIED OLIGONUCLEOTIDES

Modified oligonucleotides may be introduced into hybridization probes described in section 5.23 and used under appropriate conditions therefor. For example, pyrimidines with a halogen at the C5-position may be used to improve duplex stability by influencing base stacking.

2,6-diaminopurine may be used to provide a third hydrogen bond in base pairing with thymine, thereby thermally stabilizing DNA-duplexes. Using 2,5-diaminopurine may increase duplex stability to allow more stringent conditions for annealing, thereby improving the specificity of duplex formation, suppressing background problems and permitting the use of shorter oligomers. The synthesis of the triphosphate versions of these modified nucleotides is disclosed by Hoheisel & Lehrach(1990).

One may also use the non-discriminatory base analogue, or universal base, designed by Nichols *et. al.,* (1994). This new analogue, l-(2 -deoxy-10 -D-ribftzranosyl)-3-nitropyrrole ignated M), was generated for use in oligonucleotide probes and primers for solving the design problems that arise as a result of the degeneracy of the genetic code, or when only fragmentary peptide sequence data are available. This analogue maximizes stacking while minimizing hydrogen-bonding interactions without sterically disrupting a DNA duplex.

The M nucleoside analogue was designed to maximize stacking interactions using aprotic polar substituents linked to heteroaromatic rings, enhancing intra- and inter-stand stacking interactions to lessen the role of hydrogen bonding in base-pairing specificity. Nichols *et al.*. (1994) favored 3-nitropyrrole2 -deoxyribonucleoside because of its structural and electronic resemblance to p-nitroaniline, whose derivatives are among the smallest known intercaltors of double-stranded DNA.

The dimethoxytrityl-protectedphosphoramidite of nucleoside M is also available for incorporation into nucleotides used as primers for sequencing and polymerase chain reaction (PCR). Nichols *et. al.,* (1994) showed that a substantial number of nucleotides can be replaced by M without loss of primer specificity.

A unique property of M is its ability to replace long strings of contiguous nucleosides and still yield functional sequencing primers. Sequences with three, six and nine M substitutions have all been reported to give readable sequencing ladders, and PCR

with three different M-containing primers all resulted in amplification of the correct product (Nichols *et. al.*, 1994).

Probes may be retrieved one by one and added to subarrays covered by hybridization buffer. It is preferred that retrieved probes be placed in a new plate and labelled or mixed with hybridization buffer. The preferred method of retrieval is by accessing stored plates one by one and pipetting (or transferring by metal pins) a sufficient amount of each selected probe from each plate to specific wells in an intermediary plate. An array of individually addressable pipettes or pins may be used to speed up the retrieval process.

5.25 PREPARATION OF LABELED PROBES

The oligonucleotide probes may be prepared by automated synthesis, which is routine to those of skill in the art, for example, using and Applied Biosystems system. Alternatively, probes may be prepared using Genosys Biotechnologies Inc. Methods using stacks of porous Teflon wafers.

Oligonucleotide probes may be labeled with, for example, radioactive labels (^{35}S, ^{32}P, and preferably, ^{33}P) for arrays with 100-200 μm or 100-400 μm spots; non-radioactive isotopes (Jacobsen *et al.*. 1990); or fluorophores (Brumbaugh *et al.*. 1988). All such labeling methods are routine in the art, as exemplified by the relevant sections in Sambrook *et al.*. (1989) and by further references such as Schubert *et al.*. (1990), Murakami *et al.*. (1991) and Cate *et al.*. (1991), all articles being specifically incorporated herein by reference.

In regard to radiolabelling, the common methods are end-labeling using T4 polynucleotide kinase or high specific activity labeling using Klenow or even T7 polymerase. These are described as follows.

Synthetic oligonucleotides are synthesized without a phosphate group at their 5' termini and are therefore easily labeled by transfer of the ^{32}P or ^{33}P from [^{32}P]ATP or [^{33}P]ATP using the enzyme bacteriophage T4 polynucleotide kinase. If the reaction is carried out efficiently, the specificity activity of such probes can be as high as the specific activity of the [^{32}P]ATP or [^{33}P]ATP itself. The reaction described below is designed to label 10 pmoles of an oligonucleotide to high specific activity. Labeling of different

amounts of oligonucleotide can easily be achieved by increasing or decreasing the size of the reaction, keeping the concentrations of all components constant.

A reaction mixture would be created using 1.0 µl of oligonucleotide (10 pmoles/µl); 2.0 µl of 10 x bacteriophage T4 polynucleotide kinase buffer; 5.0 µl of [-^{32}P]ATP or [^{33}P]ATP (sp. act. 5000 Ci/mmole; 10 mCi/ml in aqueous solution) (10 pmoles); and 11.4 µl of water. Eight (8) units (10u/µl1) ofbacteriophagT4 polynucleotide kinase is added to the reaction mixture, mixed well, and incubated for 45 minutes at 37°C. The reaction is heated for 10 minutes at 68°C to inactivate the bacteriophage T4 polynucleotide kinase.

The efficiency of transfer of ^{32}P or ^{33}P to the oligonucleotide and its specific activity is then determined. If the specific activity of the probe is acceptable, it is purified. If the specific activity is too low, an additional 8 units of enzyme is added and incubated for a further 30 minutes at 37°C before heating the reaction for 10 minutes at 68°C to inactivate the enzyme.

Purification of radiolabeled oligonucleotides can be achieved by precipitation with thanol; precipitation with cetylpyridinium bromide; by chromatography through Bio-gel P-60; or by chromatography on a Sep-Pak C_{18} column (Pharmacia).

Probes of higher specific activities can be obtained using the Klenow fragment of *E. coli* DNA polymerase I to synthesize a strand of DNA complementary to the synthetic oligonucleotide. A short primer is hybridized to an oligonucleotide template whose sequence is the complement of the desired radiolabeled probe. The primer is then extended using the Klenow fragment of *E. coli* DNA polymerase I to incorporate [-^{32}P] dNTPs or [-^{33}P] dNTPs in a template-directed manner. After the reaction, the template and product are separated by denaturation followed by electrophoresis through a polyacrylamide gel under denaturing conditions. With this method, it is possible to generate oligonucleotide probes that contain several radioactive atoms per molecule of oligonucleotide if desired.

To use this method, one would mix in a microfuge tube the calculated amounts of [α-^{32}P]dNTP's or [α-^{33}P]dNTP's necessary to achieve the desired specific activity and sufficient to allow complete synthesis of all template strands. The concentration of dNTPs should not be less than IµM at any stage during the reaction. Then add to the tube

the appropriate amounts of primer and template DNAs, with the primer being in three- to tenfold molar excess over the template.

Klenow buffer is then added and mixed well. 2-4 units of the Klenow fragment of *E. coli* DNA polymerase I would then be added per 5 µl of reaction volume, mixed and incubated for 2-3 hours at 4°C. If desired, the process of the reaction may be monitored by removing small (0.1 µl) aliquots and measuring the proportion of radioactivity that has become precipitable with 10% trichloroacetic acid (TCA).

The reaction would be diluted with an equal volume of gel-loading buffer, heated to 80°C for 3 minutes, and then the entire sample loaded on a denaturing polyacrylamide gel. Following electrophoresis, the gel is autoradiographed, allowing the probe to be localized and removed from the gel. Various methods for fluorphoric labeling are also available, as follows. Brumbaugh *et al.*. (1988) describe the synthesis of fluorescently labeled primers. A deoxyuridine analog with a primary amine "linker arm" of 12 atoms attached at C-S is synthesized. Synthesis of the analog consists of derivatizing 2 - deoxyuridine through organometallic intermediates to give 5 (methyl propenoyl)-2 - deoxyuridine. Reaction dimethoxytrityl-chloride produces the corresponding 5 - dimethoxytrityl adduct. The methyl ester is hydrolyzed, activated, and reacted with an appropriately monoacylated ldiamine. After purification, the resultant linker arm nucleosides are converted to nucleoside analogs suitable for chemical oligonucleotide synthesis.

Oligonucleotides would then be made that include one or two linker arm bases by using modified phosphoridite chemistry. To a solution of 50 nmol of the linker arm oligonucleotide in 25 µl of 500 mM sodium biocarbonate (pH 9.4) is added 20 ul of 300 mM FITC in dimethyl sulfoxide. The mixture is agitated at room temperature for 6 hrs. The oligonucleotide is separated from free FITC by elution form a 1 x 30 cm Sephadex G-25 column with 20 mM ammonium acetate (pH 6), combining fractions in the first UV-absorbing peak.

In general, fluorescent labeling of an oligonucleotide at it s 5'-end initially involved two steps. First, a N-protected aminoalkyl phosphoramidite derivative is added to the 5'-end of an oligonucleotide during automated DNA synthesis. After removal of all protecting groups, the NHS ester of an appropriate fluorescent dye is coupled to the 5

-amino group overnight followed by purification of the labeled oligonucleotide from the excess of dye using reverse phase HPLC or PAGE.

Schubert *et al.*. (1990) describes the synthesis of a phosphoramidite that enables olignucleotides labeled with fluorescent tags to be produced during automated DNA synthesis.

Murakami *et al.*. also described the preparation of fluorescein-labeled oligonucleotides.

Cate *et al.*. (1991) describes the use of oligonucleotide probes directly conjugated to alkaline phosphatase in combination with a direct chemiluminescent substrate (AMPPD) to allow probe detection.

Labeled probes could readily be purchased form a variety of commercial sources, including GENSET, rather then synthesized.

Other labels include ligands which can serve as specific binding members to a labeled antibody, chemiluminescers, enzymes, antibodies which can serve as a specific binding pair member for a labeled ligand, and the like. A wide variety of labels have been employed in immunoassays which can readily be employed. Still other labels include antigens, groups with specific reactivity, and electrochemically detectable moeities.

In general, labeling of nucleic acids with electrophore mass labels ("EML") is described, for example, in Xu *et al.*., J. Chromatography 764:95-102(1997). Electrophores are compounds that can be detected with high sensitivity by electron capture mass spectrometry (EC-MS). EMLs can be attached to a probe using chemistry that is well known in the art for reversibly modifying a nucleotide (*e.g.*, well known nucleotide synthesis chemistry teaches a variety of methods for attaching molecules to nucleotides as protecting groups). EMLs are detected using a variety of well known electron capture mass spectrometry devices (*e.g.*, devices sold by Finnigan Corporation). Further, techniques that may be used in the detection of EMLs include, for example, fast atomic bombardment mass spectrometry (see, *e.g.*, Koster *et al.*., Biomedical Environ. Mass Spec. 14:111-116(1987)); plasma desorption mass spectrometry; electrospray/ionspray (see, *e.g.*, Fenn *et al.*., J. Phys. Chem. 88:4451-59(1984), PCI Appln. No. WO 90/14148, Smith cial., Anal. Chem. 62:882-89(1990)); and matrix-

assisted laser desorption/ionization (Hillenkamp, *et al..*, "Matrix Assisted UV-Laser Desorption/Ionization: A New Approach to Mass Spectrometry of Large Biomolecules," Biological Mass Spcciromctry (Burlingame and McCloskey, eds.), Elsevier Science Publishers, Amsterdam, pp. 49-60, 1990); Huth-Fehre *et al..*, "Matrix Assisted Laser Desorption Mass Spectrometry of Oligodeoxythymidylic Acids," Rapid Communications in Mass Spectrometry, 6:209-13 (1992)).

In preferred embodmients, the EMLs are attached to a probe by a covalent bond that is light sensitive. The EML is released from the probe after hybridization with a target nucleic acid by a laser or other light source emitting the desired wavelength of light. The EML is then fed into a GC-MS (gas chromatogrph-mass spectrometer) or other appropriate device, and identified by its mass.

5.26 PREPARATION OF SEQUENCING CHIPS AND ARRAYS

The nucleic acids of he present invention can be affixed to chips to create arrays. A basic example is using 6-mers attached to 50 micron surfaces to give a chip with dimensions of 3 x 3 mm which can be combined to give an array of 20 x 2020cm. Another example is using 9-mer oligonucleotides attached to 10 x 10 micron surface to create a 9-mer chip, with dimensions of 5 x 5 mm. 4000 units of such chips may be used to create a 30 x 30 cm array. In one embodiment the nucleic acids comprise an oligochip array in which 4,000 to 16,000 oligochips are arranged into a square array. A plate, or collection of tubes, as also depicted, may be packaged with the array as part of the sequencing kit.

The arrays may be separated physically from each other or by hydrophobic surfaces. One possible way to utilize the hydrophobic strip separation is to use technology such as the Iso-Grid Microbiology System produced by QA Laboratories, Toronto, Canada.

Hydrophobic grid membrane filters (HGMF) have been in use in analytical food microbiology for about a decade where they exhibit unique attractions of extended numerical range and automated counting of colonies. One commercially-available grid is Iso-GRID™ from QA Laboratories Ltd. (Toronto, Canada) which consists of a square (60 x 60 cm) of polysulfone polymer (Gelman Tuffryn HT-450, 0.4 micron pore size) on

which is printed a black hydrophobic ink grid consisting of 1600 (40 x 40) square cells. HCMF have previously been inoculated with bacterial suspensions by vacuum filtration and incubated on the differential or selective media of choice.

Because the microbial growth is confined to grid cells of known position and size on the membrane, the HGMF functions more like an MPN apparatus than a conventional plate or membrane filter. Peterkin *et al.*. (1987) reported that these HGMFs can be used to propagate and store genomic libraries when used with a HGMF replicator. One such instrument replicates growth from each of the 1600 cells of the ISO-GRID and enables many copies of the master HGMF to be made (Peterkin *et al.*., 1987).

Sharpe *et al.*. (1989) also used ISO-GRID HGMF form QA Laboratories and an automated HGMF counter (MI-I 00 Interpreter) and RP- 100 Replicator. They reported a technique for maintaining and screening many microbial cultures.

Peterkin and colleagues later described a method for screening DNA probes using the hydrophobic grid-membrane filter (Peterkin *et al.*., 1989). These authors reported methods for effective colony hybridization directly on HGMFs. Previously, poor results had been obtained due to the low DNA binding capacity of the epoxysulfone polymer on which the HGMFs are printed: However, Peterkin *et al.*. (1989) reported that the binding of DNA to the surface of the membrane was improved by treating the replicated and incubated HGMF with polyethyleneimine,a polycation, prior to contact with DNA. Although this early work uses cellular DNA attachment, and has a different objective to the present invention, the methodology described may be readily adapted for Format 3 SBH.

In order to identify useful sequences rapidly, Peterkin et al.. (1989) used radiolabeled plasmid DNA from various clones and tested its specificity against the DNA on the prepared HGMFs. In this way, DNA from recombinant plasmids was rapidly screened by colony hybridization against 100 organisms on HGMF replicates which can be easily and reproducibly prepared. Manipulation with small (2-3 mm) chips, and parallel execution of thousands of the reactions. The solution of the invention is to keep the chips and the probes in the corresponding arrays. In one example, chips containing 250,000 9-mers are synthesized on a silicon wafer in the form of 8 x 8 mM plates (15 μM/oligonucleotide, Pease *et al.*., 1994) arrayed in 8 x 12 format (96 chips) with a 1 mM groove in between. Probes

are added either by multichannel pipette or pin array, one probe on one chip. To score all 4000 6-mers, 42 chip arrays have to be used, either using different ones, or by reusing one set of chip arrays several times. In the above case, using the earlier nomenclature of the application, F=9; P=6; and F + P = 15. Chips may have probes of formula B_xN_n, where x is a number of specified bases B; and n is a number of non-specified bases, so that x = 4 to 10 and n = 1 to 4. To achieve more efficient hybridization, and to avoid potential influence of any support oligonucleotides, the specified bases can be surrounded by unspecified bases, thus represented by a formula such as $(N)_nB_x(N)_m$.

5.27 HYBRIDIZATION AND SCORING PROCESS

Labeled probes may be mixed with hybridization buffer and pipetted, preferably by multichannel pipettes, to the subarrays. To prevent mixing of the probes between subarrays (if there are no hydrophobic strips or physical barriers imprinted in the membrane), a corresponding plastic, metal or ceramic grid may be firmly pressed to the membrane. Also, the volume of the buffer may be reduced to about 1 μl or less per mm^2. The concentration of the probes and hybridization conditions used may be as described previously except that the washing buffer may be quickly poured over the array of subarrays to allow fast dilution of probes and thus prevent significant cross-hybridization. For the same reason, a minimal concentration of the probes may be used and hybridization time extended to the maximal practical level. For DNA detection and sequencing, knowledge of a "normal" sequence allows the use of the continuous stacking interaction phenomenon to increase the signal. In addition to the labelled probe, additional unlabelled probes which hybridize back to back with a labelled one may be added in the hybridization reaction. The amount of the hybrid may be increased several times. The probes may be connected by ligation. This approach may be important for resolving DNA regions forming "compressions".

In the case of radiolabelled probes, images of the filters may be obtained, preferably by phosphor storage technology. Fluorescent labels may be scored by CCD cameras, confocal microscopy or otherwise. In order to properly scale and integrate data from different hybridization experiments, raw signals are normalized based on the amount of target in each dot. Differences in the amount of target DNA per dot may be corrected for dividing signals of each probe by an average signal for all probes scored on

one dot. The normalized signals may be scaled, usually from 1-100, to compare data from different experiments. Also, in each subarray, several control DNAs may be used to determine an average background signal in those samples which do not contain a full match target. For samples obtained from diploid (polyploid) scores, homozygotic controls may be used to allow recognition of heterozygotes in the samples.

5.28 HYBRIDIZATION WITH OLIGONUCLEOTIDES

Oligonucleotides were either purchased from Genosys Inc., Houston, Texas or made on an Applied Biosystems 381A DNA synthesizer. Most of the probes used were not purified by HPLC or gel electrophoresis. For example, probes were designed to have both a single perfectly complementary target in interferon, a Ml3 clone containing a 921 bp Eco RI-Bgl II human BI-interferon fragment (Ohno and Tangiuchi, Proc. Natl. Acad. Sci. 74: 4370-4374 (1981)], and at least one target with an end base mismatch in Ml3 vector itself.

End labelling of oligonucleotides was performed as described (Maniatis *et al..*, Molccular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory Cold Spring Harbor, New York (1982)) in 10 µl containing T4-polynucleotide kinase (5 units Amersham), □^{32}P-ATP (3.3 pM, 10 mCi Amersham 3000 Ci/mM) and the oligonucleotide (4 pM, 10 ng). Specific activities of the probes were 2.5-5 x10^9 cpm/nm.

Single stranded DNA (2 to 4 ml in 0.5 NaOH, 1.5 M NaC1) was spotted on a Gene Screen membrane wetted with the same solution, the filters were neutralized in 0.05 M Na_2HPO_4 pH 6.5, baked in an oven at 80°C for 60 min, and UV irradiated for 1 min. Then, the filters were incubated in hybridization solution (0.5 M Na_2HPO_4 pH 7.2, 7% sodium lauroyl sarcosine for 5 min at room temperature and placed on the surface of a plastic Petri dish. A drop of hybridization solution(l0 ml, 0.5 M Na_2HPO_4 pH 7.2, 7% sodium lauroyl sarcosine) with a 32ρ end-labeled oligomer probe at 4 nM concentration was placed over 1-6 dots per filter, overlaid with a square piece of polyethylene (approximately 1 x 1 cm.), and incubated in a moist chamber at the indicated temperatures for 3 hr. Hybridization was stopped by placing the filter in 6X SSC washing solution for 3 to 5 minutes at 0°C to remove unhybridized probe. The filter was either dried, or further washed for the indicated times and temperatures, and

autoradiographed. For discrimination measurements, the dots were excised from the dried filters after autoradiography (a phosphoimager (Molecular Dynamics, Sunnyvale, California) may be used) placed in liquid scintillation cocktail and counted. The uncorrected ratio of cpms for IF and M13 dots is given as D.

The conditions reported herein allow hybridization with very short oligonucleotides but ensure discriminations between matched and mismatched oligonucleotides that are complementary to and therefore bind to a target nucleic acid. Factors which influence the efficient detection of hybridization of specific short sequences based on the degree of discriminations (D) between a perfectly complementary target and an imperfectly complementary target with a single mismatch in the hybrid are defined. In experimental tests, dot blot hybridization of twenty-eight probes that were 6 to 8 nucleotides in length to two M13 clones or to model oligonucleotides bound to membrane filters was accomplished. The principles guiding the experimental procedures are given below.

Oligonucleotide hybridization to filter bound target nucleic acids only a few nucleotides longer than the probe in conditions of probe excess is a pseudo-first order reaction with respect to target concentration. This reaction is defined by:

$$S_t/S_o = e^{-kh[OP]t}$$

Wherein S_t and S_o are target sequence concentrations at time t and t_0, respectively. (OP) is probe concentration and t is temperature. The rate constant for hybrid formation, k_h increases only slightly in the 0°C to 30°C range (Porschke and Eigen, *J. Mol. Biol.* 62: 361 (1971); Craig et al.. *J Mol. Biol.* 62: 383 (1971). Hybrid melting is a first order reaction with respect to hybrid concentration (here replaced by mass due to filter bound state) as shown in:

$$H_t/H_o = e^{-kmt}$$

In this equation, H_t and H_o are hybrid concentrations at times t and t_o, respectively; k_m is a rate constant for hybrid melting which is dependent on temperature and salt concentration (Ikuta et al.. *Nucl. Acids Res.* 15: 797 (1987); Porsclike and Eigen, *J. Mol. Biol.* 62: 361 (1971); Craig et al.. *J. Mol. Biol.* 62: 303 (1971)). During hybridization, which is a strand association process, the back, melting, or strand dissociation, reaction takes place as well. Thus, the amount of hybrid formed in time is

result of forward and back reactions. The equilibrium may be moved towards hybrid formation by increasing probe concentration and/or decreasing temperature. However, during washing cycles in large volumes of buffer, the melting reaction is dominant and the back reaction hybridization is insignificant, since the probe is absent. This analysis indicates workable Short Oligonucleotide Hybridization (SOH) conditions call be varied for probe concentration or temperature.

D or discrimination is defined above:

$D = H_p (t_w)/H_i (t_w)$

$H_p (t_w)$ and $H_i (t_w)$ are the amounts hybrids remaining after a washing time, t_w, for he identical amounts of perfectly and imperfectly complementary duplex, respectively. For a given temperature, the discrimination D changes with the 10 length of washing time and reaches the maximal value when $H_i = B$ which is equation five.

The background, B, represents the lowest hybridization signal detectable in the system. Since any further decrease of H_i may not be examined, D increases upon continued washing. Washing past t_w just decreases H_p relative to B, and is seen as a decrease in D. The optimal washing time, t_w, for imperfect hybrids, from equation three and equation five is:

$t_w = 1n (B / H_i (t_0)/k_{m.i}$

Since H_p is being washed for the same t_w, combining equations, one obtains the optimal discrimination function:

$D = e^{1n\ B/Hi\ (t0))\ km,p/km,i}$

The change of D as a function, of T is important because of the choice of an optimal washing temperature. It is obtained by substituting the Arrhenius equation which is:

$K- = Ae^{-Ea / RT}$

into the previous equation to form the final equation:

$D = H_p((t_0)/B\ X\ (B/H_i (t_0))^{(Ap / Ai)\ e\ (E_{a,i}-E_{a,p})/RT};$

Wherein B is less than $H_i (t_0)$.

Since the activation energy for perfect hybrids, E_{ap}, and the activation energy for imperfect hybrids, $E_{a,i}$, can be either equal, or $E_{a,i}$ less than $E_{a,p}$ D is temperature independent, or decreases with increasing temperature, respectively. This result implies

that the search for stringent temperature conditions for good discrimination in SOH is unjustified. By washing at lower temperatures, one obtains equal or better discrimination, but the time of washing exponentially increases with the decrease of temperature. Discrimination more strongly decreases with T, if $H_i(t_0)$ increases relative to H_p (t_0).

D at lower temperatures depends to a higher degree on the H_p (t_0)/B ratio than on the H_p (t_0) / H_i (t_0) ratio. This result indicates that it is better to obtain a sufficient quantity of H_p in the hybridization regardless of the discrimination that can be achieved in this step. Better discrimination can then be obtained by washing, since the higher amounts of perfect hybrid allow more time for differential melting to show an effect. Similarly, using larger amounts of target nucleic acid a necessary discrimination can be obtained even with small differences between $K_{m.p}$ and $K_{m.i}$.

Extrapolated to a more complex situation than covered in this simple model, the result is that washing at lower temperatures is even more important for obtaining discrimination in the case of hybridization of a probe having many end-mismatches within a given nucleic acid target.

Using the described theoretical principles as a guide for experiments, reliable hybridizations have been obtained with probes six to eight nucleotides in length. All experiments were performed with a floating plastic sheet providing a film of hybridization solution above the filter. This procedure allows maximal reduction in the amount of probe, and thus reduced label costs in dot blot hybridizations. The high concentration of sodium lauroyl sarcosine instead of sodium lauroyl sulfate in the phosphate hybridization buffer allows dropping the reaction from room temperature down to 12°C. Similarly, the 4-6 X SSC, 10% sodium lauroyl sarcosine buffer allows hybridization at temperatures as low as 2°C. The detergent in these buffers is for obtaining tolerable background with up to 40 nM concentrations of labelled probe. Preliminary characterization of the thermal stability of short oligonucleotide hybrids was determined on a prototype octamer with 50% G+C content, i.e. probe of sequence TGCTCATG. The theoretical expectation is that this probe is among the less stable octamers. Its transition enthalpy is similar to those of more stable heptamers or, even to probes 6 nucleotides in length (Bresslauer et al.. *Proc. Natl. Acad. Sci.* U.S.A. 83: 3746

(1986)). Parameter T_d, the temperature at which 50% of the hybrid is melted in unit time of a minute is 18°C. The result shows that T_d is 15°C lower for the 8 bp hybrid than for an 11 bp duplex (Wallace et al.. *Nucleic Acids Res.* 6: 3543 (1979)).

In addition to experiments with model oligonucleotides, an M13 vector was chosen as a system for a practical demonstration of short oligonucleotide hybridization. The main aim was to show useful end-mismatch discrimination with a target similar to the ones which will be used in various applications of the method of the invention. Oligonucleotide probes for the M13 model were chosen in such a way that the M13 vector itself contains the end mismatched base. Vector IF, an M13 recombinant containing a 921 bp human interferon gene insert, carries single perfectly matched target. Thus, IF has either the identical or a higher number of mismatched targets in comparison to the M13 vector itself. Using low temperature conditions and dot blots, sufficient differences in hybridization signals were obtained between tie dot containing the perfect and the mismatched targets and the dot containing the mismatched targets only. This was true for the 6-mer oligonucleotides and was also true for the 7 and 8-mer oligonucleotides hybridized to the large JF-M13 pair of nucleic acids.

The hybridization signal depends on the amount of target available on the filter for reaction with the probe. A necessary control is to show that the difference in sign intensity is not a reflection of varying amounts of nucleic acid in the two dots. Hybridization with a probe that has the same number and kind of targets in both IF and M13 shows that there is an equal amount of DNA in the dots. Since the efficiency of hybrid formation increases with hybrid length, the signal for a duplex having six nucleotides was best detected with a high mass of oligonucleotide target bound to the filter. Due to their lower molecular weight, a larger number of oligonucleotide target molecules can be bound to a given surface area when compared to large molecules of nucleic acid that serves as target.

To measure the sensitivity of detection with unpurified DNA, various amounts of phage supernatants were spotted on the filter and hybridized with ^{32}P-labelled octamer. As little as 50 million unpurified phage containing no more than 0.5 ng of DNA gave a detectable signal indicating that sensitivity of the short oligonucleotide hybridization method is sufficient. Reaction time is short, adding to the practicality.

As mentioned in the theoretical section above, the equilibrium yield of hybrid depends oil probe concentration and/or temperature of reaction. For instance, the signal level for the same amount of target with 4 nM octamer at 13°C is 3 times lower than with a probe concentration of 40 nM, and is decreased 4 to 5 times by raising the hybridization 5 temperature to 25°C.

The utility of the low temperature wash for achieving maximal discrimination is demonstrated. To make the phenomenon visually obvious, 50 times more DNA was put in the M13 dot than in the IF dot using hybridization with a vector specific probe. In this way, the signal after the hybridization step with the actual probe was made stronger in 10 the, mismatched that in the matched case. The H_p/H_i ratio was 1:4. Inversion of signal intensities after prolonged washing at 7°C was achieved without a massive loss of perfect hybrid, resulting in a ratio of 2:1. In contrast, it is impossible to achieve any discrimination at 25°C, since the matched target signal is already brought down to the background level with 2 minute washing; at the same time, the signal from the 15 mismatched hybrid is still detectable. The loss of discrimination at 13°C compared to 7°C is not so great but is clearly visible. If one considers the 90 minute point at 7°C and the 15 minute point at 13°C when, the mismatched hybrid signal is near the background level, which represents optimal washing times for the respective conditions, it is obvious that the amount of several times greater at 7°C than at 13°C. To illustrate this further, the 20 time course of the change discrimination with washing of the same amount of starting hybrid at the two temperatures shows the higher maximal D at the lower temperature. These results confirm the trend in the change of D with temperature and the ratio of amounts of the two types of hybrid at the start of the washing step.

In order to show the general utility of the short oligonucleotide hybridization 25 conditions, we have looked hybridization of 4 heptamers, 10 octamers and an additional 14 probes up to 12 nucleotides in length in our simple M13 system. These include-the nonamer GTTTTTTAA and octamer GGCAGGCG representing the two extremes of GC content. Although GC content and sequence are expected to influence the stability of short hybrids (Bresslauer et al.. *Proc. Natl. Acad. Sci.* U.S.A. 83: 3746 (1986)), the low 30 temperature short oligonucleotide conditions were applicable to all tested probes in achieving sufficient discrimination. Since the best discrimination value obtained with

probes 13 nucleotides in length was 20, a several fold drop due to sequence variation is easily tolerated.

The M13 system has the advantage of showing the effects of target DNA complexity on the levels of discrimination. For two octamers having either none or five mismatched targets and differing in only one GC pair the observed discriminations were 18.3 and 1.7, respectively.

In order to show the utility of this method, three probes 8 nucleotides in length were tested on a collection of 51 plasmid DNA dots made from a library in Bluescript vector. One probe was present and specific for Bluescript vector but was absent in M13, while the other two probes had targets that were inserts of known sequence. This system allowed the use of hybridization negative or positive control DNAs with each probe. This probe sequence (CTCCCTTT) also had a complementary target in the interferon insert. Since the M13 dot is negative while the interferon insert in either M13 or Bluescript was positive, hybridization is sequence specific. Similarly, probes that detect the target sequence in only one of 51 inserts, or in none of the examined inserts along with controls that confirm that hybridization would have occurred if the appropriate targets were present in the clones.

Thermal stability curves for very short oligonucleotide hybrids that are 6-8 nucleotides in length are at least 15°C lower than for hybrids 11-12 nucleotides in length (Wallace *et al.*. *Nucleic Acids Res.* 6: 3543-3557 (1979). However, performing the hybridization reaction at a low temperature and with a very practical 0.4-40 nM concentration of oligonucleotide probe allows the detection of complementary sequence in known or unknown nucleic acid target. To determine an unknown nucleic acid sequence completely, an entire set containing 65,535 8-mer probes may be used. Sufficient amounts of nucleic acid for this purpose are present in convenient biological samples such as a few microliters of M13 culture, a plasmid prep from 10 ml of bacterial culture or a single colony of bacteria, or less than 1 ml of a standard PCR reaction.

Short oligonucleotides 6-10 nucleotides long give excellent discrimination. The relative decrease in hybrid stability with a single end mismatch is greater than for longer probes. Results with the octamer TGCTCATG support this conclusion. In the experiments, the target with a G/T end mismatch, hybridization to the target of this type

of mismatch is the most stable of all other types of oligonucleotide. This discrimination achieved is the same as or greater than an internal G/T mismatch in a 19 base paired duplex greater than an internal G/T mismatch in a 19 paired duplex *see,* (Ikuta *et al.*. Nucl. Acids Res. 15: 797 (1987). Exploiting these discrimination properties using the described hybridization conditions for short oligonucleotide hybridization allows a very precise determination of oligonucleotide targets. In contrast to the ease of detecting discrimination between perfect and imperfect hybrids, a problem that may exist with using very short oligonucleotides is the preparation of sufficient amounts of hybrids. In practice, the need to discriminate H_p and H_i is aided by increasing the amount of DNA in the dot and/or the probe concentration, or by decreasing the hybridization temperature. However, higher probe concentrations usually increase background. Moreover, there are limits to the amounts of target nucleic acid that are practical to use. This problems was solved by the higher concentration of the detergent Sarcosyl which gave an effective background with 4 nM of probe. Further improvements may be effected either in the use of competitors for unspecific binding of probe to filter, or by changing the hybridization support material. Moreover, for probes having E_a less than 45 Kcal/mol (*eg.*, for many heptamers and a majority of hexamers, modified oligonucleotides give a more stable hybrid (Asseline, et al.. *Proc. Nat'l Acad. Sci.* 81: 3297 (1984) than their unmodified counterparts. The hybridization conditions described in this invention for short oligonucleotide hybridization using low temperatures give better discriminating for all sequences and duplex hybrid inputs. The only price paid in achieving uniformity in hybridization conditions for different sequences is an increase in washing time from minutes to up to 24 hours depending on the sequence. Moreover, the washing time can be further reduced by decreasing the salt concentration.

Although there is excellent discrimination of one matched hybrid over a mismatched hybrids, in short oligonucleotide hybridization, signals from mismatched hybrids exist, with the majority of the mismatch hybrids resulting from end mismatch. This may limit insert sizes that may be effectively examined by a probe of a certain length.

The influence of sequence complexity on discrimination cannot be ignored. However, the complexity effects are more significant when defining sequence

information by short oligonucleotide hybridization for specific, nonrandom sequences, and can be overcome by using an appropriate probe to target length ratio. The length ratio is chosen to make unlikely, on statistical grounds, the occurrence of specific sequences which have a number of end-mismatches which would be able to eliminate or falsely invert discrimination. Results suggest the use of oligonucleotides 6, 7, and 8 nucleotides in length on target nucleic acid inserts shorter than 0.6, 2.5, and 10 kb, respectively.

5.29 DNA SEQUENCING

An array of subarrays allows for efficient sequencing of a small set of samples arrayed in the form of replicated subarrays; for example, 64 samples may be arrayed on a 8 x 8 mm subarray and 16 x 24 subarrays may be replicated on a 15 x 23 cm membrane with 1 mm wide spacers between the subarrays. Several replica membranes may be made. For example, probes from a universal set of three thousand seventy-two 7-mers may be divided in thirty-two 96-well plates and labelled by kinasing. Four membranes may be processed in parallel during one hybridization cycle. On each membrane, 384 probes may be scored. All probes may be scored in two hybridization cycles. Hybridization intensities may be scored and the sequence assembled as described below.

If a single sample subarray or subarrays contains several unknowns, especially when similar samples are used, a smaller number of probes may be sufficient if they are intelligently selected on the basis of results of previously scored probes. For example, if probe AAAAAAA is not positive, there is a small chance that any of 8 overlapping probes are positive. If AAAAAAA is positive, then two probes are usually positive. The sequencing process in this case consists of first hybridizing a subset of minimally overlapped probes to define positive anchors and then to successively select probes which confirms one of the most likely hypotheses about the order of anchors and size and type of gaps between them. In this second phase, pools of 2-10 probes may be used where each probe is selected to be positive in only one DNA sample which is different from the samples expected to be positive with other probes from the pool.

The subarray approach allows efficient implementation of probe competition (overlapped probes) or probe cooperation (continuous stacking of probes) in solving

branching problems. After hybridization of a universal set of probes the sequence assembly program determines candidate sequence subfragments (SFs). For the further assembly of SFs, additional information has to be provided (from overlapped sequences of DNA fragments, similar sequences, single pass gel sequences, or from other hybridization or restriction mapping data). Competitive hybridization and continuous stacking interactions have been proposed for SF assembly. These approaches are of limited practical value for sequencing of large numbers of samples by SBH wherein a labelled probe is applied to a sample affixed to an array if a uniform array is used. Fortunately, analysis of small numbers of samples using replica subarrays allows efficient implementation of both approaches. On each of the replica subarrays, one branching point may be tested for one or more DNA samples using pools of probes similarly as in solving mutated sequences in different samples spotted in the same subarray (see above).

If in each of 64 samples described in this example, there are about 100 branching points, and if 8 samples are analyzed in parallel in each subarray, then at least 800 subarray probings solve all branches. This means that for the 3072 basic probings an additional 800 probings (25%) are employed. More preferably, two probings are used for one branching point. If the subarrays are smaller, less additional probings are used. For example, if subarrays consist of 16 samples, 200 additional probings may be scored (6%). By using 7-mer probes ($N_{1-2}B_7N_{1-2}$) and competitive or collaborative branching solving approaches or both, fragments of about 1000 bp fragments may be assembled by about 4000 probings. Furthermore, using 8-mer probes (NB_8N) 4 kb or longer fragments may be assembled with 12,000 probings. Gapped probes, for example, NB_4NB_3N or NB_4NB_4N may be used to reduce the number of branching points.

5.30 DNA ANALYSIS BY TRANSIENT ATTACHMENT TO SUBARRAYS OF PROBES AND LIGATION OF LABELLED PROBES

Oligonucleotide probes having an informative length of four to 40 bases are synthesized by standard chemistry and stored in tubes or in multiwell plates. Specific sets of probes comprising one to 10,000 probes are arrayed by deposition or *in situ* synthesis on separate supports or distinct sections of a larger support. In the last case, sections or subarrays may be separated by physicalor hydrophobic barriers. The probe

arrays may be prepared by *in situ* synthesis. A sample DNA of appropriate size is hybridized with one or more specific arrays. Many samples may be interrogated as pools at the same subarrays or independently with different subarrays within one support. Simultaneously with the sample or subsequently, a single labelled probe or a pool of labelled probes is added on each of the subarrays. If attached and labelled probes hybridize back to back on the complementary target in the sample DNA they are ligated. Occurrence of ligation will be measured by detecting a label from the probe.

This procedure is a variant of the described DNA analysis process in which DNA samples are not permanently attached to the support. Transient attachment is provided by probes fixed to the support. In this case there is no need for a target DNA arraying process. In addition, ligation allows detection of longer oligonucleotide sequences by combining short labelled probes with short fixed probes.

The process has several unique features. Basically, the transient attachment of the target allows its reuse. After ligation occur the target may be released and the label will stay covalently attached to the support. This feature allows cycling the target and production of detectable signal with a small quantity of the target. Under optimal conditions, targets do not need to be amplified, e.g. natural sources of the DNA samples may be directly used for diagnostics and sequencing purposes. Targets may be released by cycling the temperature between efficient hybridization and efficient melting of duplexes. More preferably, there is no cycling. The temperature and concentrations of components may be defined to have an equilibrium between free targets and targets entered in hybrids at about 50:50% level. In this case there is a continuous production of ligated products. For different purposes different equilibrium ratios are optimal.

An electric field may be used to enhance target use. At the beginning, a horizontal field pulsing within each subarray may be employed to provide for faster target sorting. In this phase, the equilibrium is moved toward hybrid formation, and unlabelled probes may be used. After a target sorting phase, an appropriate washing (which may be helped by a vertical electric field for restricting movement of the samples) may be performed. Several cycles of discriminative hybrid melting, target harvesting by hybridization and ligation and removing of unused targets may be introduced to increase specificity. In the next step, labelled probes are added and vertical electrical pulses may

be applied. By increasing temperature, an optimal free and hybridized target ratio may be achieved. The vertical electric field prevents diffusion of the sorted targets.

The subarrays of fixed probes and sets of labelled probes (specially designed or selected from a universal probe set) may be arranged in various ways to allow an efficient and flexible sequencing and diagnostics process. For example, if a short fragment (about 100-500 bp) of a bacterial genome is to be partially or completely sequenced, small arrays of probes (5-30 bases in length) designed on the bases of known sequence may be used. If interrogated with a different pool of 10 labelled probes per subarray, an array of 10 subarrays each having 10 probes, allows checking of 200 bases, assuming that only two bases connected by ligation are scored. Under the conditions where mismatches are discriminated throughout the hybrid, probes may be displaced by more than one base to cover the longer target with the same number of probes. By using long probes, the target may be interrogated directly without amplification or isolation from the rest of DNA in the sample. Also, several targets may be analyzed (screened for) in one sample simultaneously. If the obtained results indicate occurrence of a mutation (or a pathogen), additional pools of probes may be used to detect type of the mutation or subtype of pathogen. This is a desirable feature of the process which may be very cost effective in preventive diagnosis where only a small fraction of patients is expected to have an infection or mutation.

In the processes described in the examples, various detection methods may be used, for example, radiolabels, fluorescent labels, enzymes or antibodies (chemiluminescence), large molecules or particles detectable by light scattering or interferometric procedures.

5.31 SEQUENCING A TARGET USING OCTAMERS AND NONAMERS

Data resulting from the hybridization of octamer and nonamer oligonucleotides shows that sequencing by hybridization provides an extremely high degree of accuracy. In this experiment, a known sequence was used to predict a series of contiguous overlapping component octamer and nonamer oligonucleotides.

In addition to the perfectly matching oligonucleotides, mismatch oligonucleotides, mismatch oligonucleotides wherein internal or end mismatches occur in the duplex

formed by the oligonucleotide and the target were examined. In these analyses, the lowest practical temperature was used to maximize hybridization formation. Washes were accomplished at the same or lower temperatures to ensure maximal discrimination by utilizing the greater dissociation rate of mismatch versus matched oligonucleotide/target hybridization. These conditions are shown to be applicable to all sequences although the absolute hybridization yield is shown to be sequence dependent.

The least destabilizing mismatch that can be postulated is a simple end mismatch, so that the test of sequencing by hybridization is the ability to discriminate perfectly matched oligonucleotide/targetduplexes from end-mismatched oligonucleotide/targetduplexes.

The discriminative values for 102 of 105 hybridizing oligonucleotides in a dot blot format were greater than 2 allowing a highly accurate generation of the sequence. This system also allowed an analysis of the effect of sequence on hybridization formation and hybridization instability.

One hundred base pairs of a known portion of a human-interferon genes prepared by PCR, i.e. a 100 bp target sequence, was generated with data resulting from the hybridization of 105 oligonucleotides probes of known sequence to the target nucleic acid. The oligonucleotide probes used included 72 octamer and 21 nonamer oligonucleotides whose sequence was perfectly complementary to the target. The set of 93 probes provided consecutive overlapping frames of the target sequence e displaced by one or two bases.

To evaluate the effect of mismatches, hybridization was examined for 12 additional probes that contained at least one end mismatch when hybridized to the 100 bp test target sequence. Also tested was the hybridization of twelve probes with target end-mismatched to four other control nucleic acid sequences chosen so that the 12 oligonucleotides formed perfectly matched duplex hybrids with the four control DNAs. Thus, the hybridization of internal mismatched, end-mismatched and perfectly matched duplex pairs of oligonucleotide and target were evaluated for each oligonucleotide used in the experiment. The effect of absolute DNA target concentration on the hybridization with the test octamer and nonamer oligonucleotides was determined by defining target

DNA concentration by detecting hybridization of a different oligonucleotide probe to a single occurrence non-target site within the co-amplified plasmid DNA.

The results of this experiment showed that all oligonucleotides containing perfect matching complementary sequence to the target or control DNA hybridized more strongly than those oligonucleotides having mismatches. To come to this conclusion, we examined H_p and D values for each probe. H_p defines the amount of hybrid duplex formed between a test target and an oligonucleotide probe. By assigning values of between 0 and 10 to the hybridization obtained for the 105 probes, it was apparent that 68.5% of the 105 probes had an H_p greater than 2.

Discrimination (D) values were obtained where D was defined as the ratio of signal intensities between 1) the dot containing a perfect matched duplex formed between test oligonucleotide and target or control nucleic acid and 2) the dot containing a mismatch duplex formed between the same oligonucleotide and a different site within the target or control nucleic acid. Variations in the value of D result from either 1) perturbations in the hybridization efficiency which allows visualization of signal over background, or 2) the type of mismatch found between the test oligonucleotide and the target. The D values obtained in this experiment were between 2 and 40 for 102 of the 105 oligonucleotide probes examined. Calculations of D for the group of 102 oligonucleotides as a whole showed the average D was 10.6.

There were 20 cases where oligonucleotide/targetduplexes exhibited an end-mismatch. In five of these, D was greater than 10. The large D value in these cases is most likely due to hybridization destabilization caused by other than the most stable (G/T and G/A) end mismatches. The other possibility is there was an error in the sequence of either the oligonucleotides or the target.

Error in the target for probes with low H_p, was excluded as a possibility because such an error would have affected the hybridization of each of the other eight overlapping oligonucleotides. There was no apparent instability due to sequence mismatch for the other overlapping oligonucleotides, indicating the target sequence was correct. Error in the oligonucleotide sequence was excluded as a possibility after the hybridization of seven newly synthesized oligonucleotides was re-examined. Only 1 of the seven oligonucleotides resulted in a better D value. Low hybrid formation values may result

from hybrid instability or from an inability to form hybrid duplex. An inability to form hybrid duplexes would result from either 1) self complementarity of the chosen probe or 2) target/target self hybridization. Oligonucleotide/oligonucleotideduplex formation may be favored over oligonucleotide/targethybrid duplex formation if the probe was self-complementary. Similarly, target/target association may be favored if the target was self-complementary or may form internal palindromes. In evaluating these possibilities, it was apparent from probe analysis that the questionable probes did not form hybrids with themselves. Moreover, in examining the contribution of target/target hybridization, it was determined that one of the questionable oligonucleotide probes hybridized inefficiently with two different DNAs containing the same target. The low probability that two different DNAs have a self-complementary region for the same target sequence leads to the conclusion that target/target hybridization did not contribute to low hybridization formation. Thus, these results indicate that hybrid instability and not the inability to form hybrids was the cause of the low hybrid formation observed for specific oligonucleotides. The results also indicate that low hybrid formation is due to the specific sequences of certain oligonucleotides. Moreover, the results indicate that reliable results may be obtained to generate sequences if octamer and nonamer oligonucleotides are used.

These results show that using the methods described long sequences of any target nucleic acid may be generated by maximal and unique overlap of constituent oligonucleotides. Such sequencing methods are dependent on the content of the individual component oligomers regardless of their frequency and their position.

The sequence which is generated using the algorithm described below is of high fidelity. The algorithm tolerates false positive signals from the hybridization dots as is indicated from the fact the sequence generated from the 105 hybridization values, which included four less reliable values, was correct. This fidelity in sequencing by hybridization is due to the "all or none" kinetics of short oligonucleotide hybridization and the difference in duplex stability that exists between perfectly matched duplexes and mismatched duplexes. The ratio of duplex stability of matched and end-mismatched duplexes increases with decreasing duplex length. Moreover, binding energy decreases with decreasing duplex length resulting in a lower hybridization efficiency. However, the

results provided show that octamer hybridization allows the balancing of the factors affecting duplex stability and discrimination to produce a highly accurate method of sequencing by hybridization. Results presented in other examples show that oligonucleotides that are 6, 7, or 8 nucleotides can be effectively used to generate reliable sequence on targets that are 0.5 kb (for hexamers) 2 kb (for septamers) and 6kb (for octamers). The sequence of long fragments may be overlapped to generate a complete genome sequence.

5.32 ANALYZING THE DATA OBTAINED

Image files are analyzed by an image analysis program, like DOTS program (Drmanac *et al.*. 1993), and scaled and evaluated by statistical functions included, *e.g.,* in SCORES program (Drmanac *et al.*. 1994). From the distribution of the signals an optimal threshold is determined for transforming signal into +/- output. From the position of the label detected, F + P nucleotide sequences from the fragments would be determined by combining the known sequences of the immobilized and labeled probes corresponding to the labeled positions. The complete nucleic acid sequence or sequence subfragments of the original molecule, such as a human chromosome, would then be assembled from the overlapping F + P sequence determined by computational deduction.

One option is to transform hybridization signals *e.g.,* scores, into +/- output during the sequence assembly process. In this case, assembly will start with a F + P sequence with a very high score, for example F + P sequence AAAAAATTTTTT. Scores of all four possible overlapping probes AAAAATTTTTTA, AAAAATTTTTT, AAAAATTTTTTC and AAAAATTTTTTG and three additional probes that are different at the beginning (TAAAAATTTTTT, CAAAAATTTTTT, GAAAAATTTTTT, are compared and three outcomes defined: (i) only the starting probe and only one of the four overlapping proves have scores that are significantly positive relatively to the other six probes, in this case the AAAAAATTTTTT sequence will be extended for one nucleotide to the right; (ii) no one probe except the starting probe has a significantly positive score, assembly will stop, *e.g.,* the AAAAAATTTTT sequence is at the end of the DNA molecule that is sequenced; (iii) more than one significantly positive probe

among the overlapped and/or other three probes is found; assembly is stopped because of the error or branching (Drmanac *et al..*, 1989).

The processes of computational deduction would employ computer programs using existing algorithms (see, *e.g.,* Pevzner, 1989; Drmanac *et al..*, 1991; Labat and Drmanac, 1993; each incorporated herein by reference).

If, in addition to F + P, F (1)P, F (2)P, F(3)P or F(4)P are determined, algorithms will be used to match all data sets to correct potential errors or to solve the situation where there is a branching problem (see, *e.g.,* Drmanac *et al..*, 1989; Bains *et al..*, 1988; each incorporated herein by reference).

5.33 CONDUCTING SEQUENCING BY TWO STEP HYBRIDIZATION

Sequencing is accomplished as described herein. First, the whole chip is hybridized to a mixture of DNA as complex as 100 million of bp (one human chromosome). Guidelines for conducting hybridization can be found in papers such as Drmanac *et al..* (1990); Khrapko *et al..* (1991); and Broude *et al..* (1994). These articles teach the ranges of hybridization temperatures, buffers and washing steps that are appropriate for use in the initial steps of Format 3 SBH.

The present invention particularly contemplates that hybridization is to be carried out for up to several hours in high salt concentrations at a low temperature (-2°C to 5°C) because of a relatively low concentration of target DNA that can be provided. For this purpose, SSC buffer is used instead of sodium phosphate buffer (Drmanac *et al..*, 1990), which precipitates at 10°C. Washing does not have to be extensive (a few minutes) because of the second step, and can be completely eliminated when the hybridization cycling is used for the sequencing of highly complex DNA samples. The same buffer is used for hybridization and washing steps to be able to continue with the second hybridization step with labeled probes.

After proper washing using a simple robotic device on each array, *e.g.,* a 8 x 8 mm array, one labeled, probe, *e.g.,* a 6-mer, would be added. A 96-tip or 96-pin device would be used, performing this in 42 operations. Again, a range of discriminatory conditions could be employed, as previously described in the scientific literature.

The present invention particularly contemplates the use of the following conditions. First, after adding labeled probes and incubating for several minutes only (because of the high concentration of added oligonucleotides) at a low temperature (0-5°C), the temperature is increased to 3-10°C, depending on F + P length, and the washing buffer is added. At this time, the washing buffer used is one compatible with any ligation reaction (*e.g.*, 100 mM salt concentration range). After adding ligase, the temperate is increased again to 15-37°C to allow fast ligation (less than 30 min) and further discrimination of full match and mismatch hybrids.

The use of cationic detergents is also contemplated for use in Format 3 SBH, as described by Pontius & Berg (1991, incorporated herein by reference). These authors describe the use of two simple cationic detergents, dodecy- and cetyltrimethylamnionium bromide (DTAB and CTAB) in DNA renaturation.

DTAB and CTAB are variants of the quaternary amine tetramethylammonium bromide (TMAB) in which one of the methyl groups is replaced by either a 12-carbon (DTAB) or a 16-carbon (CTAB) alkyl group. TMAB is the bromide salt of the tetramethylammoniumcation, a reagent used in nucleic acid renaturation experiments to decrease the G-C content bias of the melting temperature. DTAB and CTAB are similar in structure to sodium dodecyl sulfate (SDS), with the replacement of the negatively charged sulfate of SDS by a positively charged quaternary amine. While SDS is commonly used in hybridization buffers to reduce nonspecific binding and inhibit nucleases, it does not greatly affect the rate of renaturation.

When using a ligation process, the enzyme could be added with the labeled probes or after the proper washing step to reduce the background. Although not previously proposed for use in any SBH method, ligase technology is well established within the field of molecular biology. For example, Hood and colleagues described a ligase-mediated gene detection technique (Landegren *et al..*, 1988), the methodology of which can be readily adapted for use in Format 3 SBH. Wu & Wallace also describe the use of bacteriophage T4 DNA ligase to join two adjacent, short synthetic olignucleotides. Their oligoligation reactions were carried out in 50 mM Tris HCl pH 7.6, 10 mM $MgCl_2$, 1 mM ATP, 1 mM DTT, and 5% PEG. Ligation reactions were heated to 100°C for 5-10 min followed by cooling to 0°C prior to the addition of T4 DNA ligase (1 unit; Bethesda

Research Laboratory, gaithersburg, MD). Most ligation reactions were carried out at 30°C and terminated by heating to 100°C for 5 min.

Final washing appropriate for discriminating detection of hybridized adjacent, or ligated, oligonucleotidesof length (F + P), is then performed. This washing step is done in water for several minutes at 40-60°C to wash out all the non-ligated labeled probes, and all other compounds, to maximally reduce background. Because of the covalently bound labeled oligonucleotides, detection is simplified (it does not have time and low temperature constraints).

Depending on the label used, imaging of the chips is done with different apparati. For radioactive labels, phosphor storage screen technology and PhosphorImager as a scanner may be used (Molecular Dynamics, Sunnyvale, CA). Chips are put in a cassette and covered by a phosphorous screen. After 1-4 hours of exposure, the screen is scanned and the image file stored at a computer hard disc. For the detection of fluorescent labels, CCD cameras and epifluorescent or confocal microscopy are used. For the chips generated directly on the pixels of a CCD camera, detection can be performed as described by Eggers *et al.*. (1994, incorporated herein by reference).

Charge-coupled device (CCD) detectors serve as active solid supports that quantitatively detect and image the distribution of labeled target molecules in probe-based assays. These devices use the inherent characteristics of microelectronics that accommodate highly parallel assays, ultrasensitive detection, high throughput, integrated data acquisition and computation. Eggers *et al.*. (1994) describe CCDs for use with probe-based assays, such as Format 3 SBH of the present invention, that allow quantitative assessment within seconds due to the high sensitivity and direct coupling employed.

The integrated CCD detection approach enables the detection of molecular binding events on chips. The detector rapidly generates a two-dimensional pattern that uniquely characterizes the sample. In the specific operation of the CCD-based molecular detector, distinct biological probes are immobilized directly on the pixels of a CCD or can be attached to a disposable cover slip placed on the CCD surface. The sample molecules can be labeled with radioisotope, chemiluminescent or fluorescent tags.

Upon exposure of the sample to the CCD-based probe array, photons or radioisotope decay products are emitted at the pixel locations where the sample has bound, in the case of Format 3, to two complementary probes. In turn, electron-hole pairs are generated in the silicon when the charged particles, or radiation from the labeled sample, are incident on the CCD gates. Electrons are then collected beneath adjacent CCD gates and sequentially read out on a display module. The number of photoelectrons generated at each pixel is directly proportional to the number of molecular binding events in such proximity. Consequently, molecular binding can be quantitatively determined (Eggers *et al.*. 1994).

By placing the imaging array in proximity to the sample, the collection efficiency is improved by a factor of at least 10 over lens-based techniques such as those found in conventional CCD cameras. That is, the sample (emitter) is in near contact with the detector (imaging array). and this eliminates conventional imaging optics such as lenses and mirrors.

When radioisotopes are attached as reporter groups to the target molecules, energetic particles are detected. Several reporter groups that emit particles of varying energies have been successfully utilized with the micro-fabricated detectors, including ^{32}P, ^{33}P, ^{35}P, ^{14}C and ^{125}I. The higher energy particles, such as from ^{32}P, provide the highest molecular detection sensitivity, whereas the lower energy particles, such as from ^{35}S, provide better resolution. Hence the choice of the radioisotope reporter can be tailored as required. Once the particular radioisotope label is selected, the detection performance can be predicted by calculating the signal-to-noise ration (SNR), as described by Eggers *et al.*. (1994).

An alternative luminescent detection procedure involves the use of fluorescent or chemiluminescent reporter groups attached to the target molecules. The fluorescent labels can be attached covalently or through interaction. Fluorescent dyes, such as ethidium bromide, with intense absorption bands in the near UV (300-350 nm) range and principal emission bands in the visible (500-650 nm) range, are most suited for the CCD devices employed since the quantum efficiency is several orders of magnitude lower at the excitation wavelength then at the fluorescent signal wavelength.

From the perspective of detecting luminescence, the polysilicon CCD gates have in capacity to filter away the contribution of incident light in the UV range, yet are very sensitive to the visible luminescence generated by the fluorescent reporter groups. Such inherently large discrimination against UV excitation enables large SNRs (greater than 100) to be achieved by the CCDs as formulated in the incorporated paper by Eggers *et al.*. (1994).

For probe immobilization on the detector, hybridization matrices may be produced on inexpensive SiO_2 wafers, which are subsequently placed on the surface of the CCD following hybridization and drying. This format is economically efficient since the hybridization of the DNA is conducted on inexpensive disposable SiO_2 wafers, thus allowing reuse of the more expensive CCD detector. Alternatively, the probes can be immobilized directly on the CCD to create a dedicated probe matrix.

To immobilize probes upon the SiO_2 coating, a uniform epoxide layer is linked to the film surface, employing an epoxy-silane reagent and standard SiO_2 modification chemistry. Amine-modified oligonucleotide probes are then linked to the SiO_2 surface by means of secondary amine formation with the epoxide ring. The resulting linkage provides 17 rotatable bonds of separation between the 3 base of the oligonucleotide and the SiO_2 surface. To ensure complete amine deprotonation and to minimize secondary structure formation during coupling, the reaction is performed in 0.1 M KOH and incubated at 37°C for 6 hours.

In Format 3 SBH in general, signals are scored per each of billion points. It would not be necessary to hybridize all arrays, *e.g.*, 4000 5 x 5 mm, at a time and the successive use of smaller number of arrays is possible.

Cycling hybridizations are one possible method for increasing the hybridization signal. In one cycle, most of the fixed probes will hybridize with DNA fragments with tail sequences non-complementary for labeled probes. By increasing the temperature, those hybrids will be melted. In the next cycle, some of them (~0.1%) will hybridize with an appropriate DNA fragment and additional labeled probes will be ligated. In this case, there occurs a discriminative melting of DNA hybrids with mismatches for both probe sets simultaneously.

In the cycle hybridization, all components are added before the cycling starts, at the T4, or a higher temperature for a thermostable ligase. Then the temperature is decreased to 15-37°C and the chip is incubated for up to 10 minutes, and then the temperature is increased to 37°C or higher for a few minutes and then again reduced. Cycles can be repeated up to 10 times. In one variant, an optimal higher temperature (10-50°C) can be used without cycling and longer ligation reaction can be performed (1-3 hours).

The procedure described herein allows complex chip manufacturing using standard synthesis and precise spotting of oligonucleotides because a relatively small number of oligonucleotides are necessary. For example, if all 7-mer oligos are synthesized (16384 probes), lists of 256 million 14-mers can be determined.

One important variant of the invented method is to use more than one differently labeled probe per base array. This can be executed with two purposes in mind; multiplexing to reduce number of separately hybridized arrays; or to determine a list of even longer oligosequences such as 3 x 6 or 3 x 7. In this case, if two labels are used, the specificity of the 3 consecutive oligonucleotides can be almost absolute because positive sites must have enough signals of both labels.

A further and additional variant is to use chips containing BxNy probes with y being from 1 to 4. Those chips allow sequence reading in different frames. This can also be achieved by using appropriate sets of labeled probes or both F and P probes could have some unspecified end positions (*i.e.*, some element of terminal degeneracy). Universal bases may also be employed as part of a linker to join the probes of defined sequence to the solid support. This makes the probe more available to hybridization and makes the construct more stable. If a probe has 5 bases, one may, *e.g.*, use 3 universal bases as a linker.

5.34 DETERMINING SEQUENCE FROM HYBRIDIZATION DATA

Sequence assembly may be interrupted where ever a given overlapping (N-1) mer is duplicated two or more times. Then either of the two N-mers differing in the last nucleotide may be used in extending the sequence. This branching point limits unambiguous assembly of sequence.

Reassembling the sequence of known oligonucleotides that hybridize to the target nucleic acid to generate the complete sequence of the target nucleic acid may not be accomplished in some cases. This is because some information may be lost if the target nucleic acid is not in fragments of appropriate size in relation to the size of oligonucleotide that is used for hybridizing. The quantity of information lost is proportional to the length of a target being sequenced. However, if sufficiently short targets are used, their sequence may be unambiguously determined.

The probable frequency of duplicated sequences that would interfere with sequence y which is distributed along a certain length of DNA may be calculated. This derivation requires the introduction of the definition of a parameter having to do with sequence organization: the sequence subfragment (SF). A sequence subfragment results if any part of the sequence of a target nucleic acid starts and ends with an (N-1) mer that is repeated two or more times within the target sequence. Thus, subfragments are sequences generated between two points of branching in the process of assembly of the sequences in the method of the invention. The sum of all subfragmentsis longer than the actual target nucleic acid because of overlapping short ends. Generally, subfragments may not be assembled in a linear order without additional information since they have shared (N-1) mers at their ends and starts. Different numbers of subfragments are obtained for each nucleic acid target depending on the number of its repeated (N-1) mers. The number depends on the value of N-1 and the length of the target.

Probability calculations can estimate the interrelationshipof the two factors. If the ordering of positive N-mers is accomplished by using overlapping sequences of length N-1 or at an average distance of A_0, the N-1 of a fragment L_f bases long is given by equation one:

$$N_{sf} = 1+A_o \text{ X } KXP\ (K, L_f)$$

Where K greater than or = 2, and P (K, L_f) represents the probability of an N-mer occurring K-times on a fragment L_f base long. Also, a computer program that is able to form subfragments from the content of N-mers for any given sequence is described below.

The number of subfragments increases with the increase of lengths of fragments for a given length of probe. Obtained subfragments may not be uniquely ordered among

themselves. Although not complete, this information is very useful for comparative sequence analysis and the recognition of functional sequence characteristics. This type of information may be called partial sequence. Another way of obtaining partial sequence is the use of only a subset of oligonucleotide probes of a given length.

There may be relatively good agreement between predicted sequence according to theory and a computer simulation for a random DNA sequence. For instance, for N-1 = 7, [using an 8-mer or groups of sixteen 10-mers of type 5' (A,T,C,G) B_8 (A,T,C,G) 3'] a target nucleic acid of 200 bases will have an average of three subfragments. However, because of the dispersion around the mean, a library of target nucleic acid should have inserts of 500 bp so that less than 1 in 2000 targets have more than three subfragments. Thus, in an ideal case of sequence determination of a long nucleic acid of random sequence, a representative library with sufficiently short inserts of target nucleic acid may be used. For such inserts, it is possible to reconstruct the individual target by the method of the invention. The entire sequence of a large nucleic acid is then obtained by overlapping of the defined individual insert sequences.

To reduce the need for very short fragments, e.g. 50 bases for 8-mer probes. The information contained in the overlapped fragments present in every random DNA fragmentation process like cloning, or random PCR is used. It is also possible to use pools of short physical nucleic acid fragments. Using 8-mers or 11-mers like 5' (A, T, C, G) N_8 (A, T, C ,G) 3' for sequencing 1 megabase, instead of needing 20,000 50 bp fragments only 2,100 samples are sufficient. This number consists of 700 random 7 kb clones (basic library), 1250 pools of 20 clones of 500 bp (subfragments ordering library) and 150 clones from jumping (or similar) library. The developed algorithm (see Example 1 8) regenerates sequence using hybridization data of these described samples.

5.35 ALGORITHM

This example describes an algorithm for generation of a long sequence written in a four letter alphabet from constituent k-tuple words in a minimal number of separate, randomly defined fragments of a starting nucleic acid sequence where K is the length of an oligonucleotide probe. The algorithm is primarily intended for use in the sequencing by hybridization (SBH) process. The algorithm is based on subfragments (SF),

informative fragments (IF) and the possibility of using pools of physical nucleic sequences for defining informative fragments.

As described, subfragments may be caused by branch points in the assembly process resulting from the repetition of a K-l oligomer sequence in a target nucleic acid. Subfragments are sequence fragments found between any two repetitive words of the length K-1 that occur in a sequence. Multiple occurrences of K-1 words are the cause of interruption of ordering the overlap of K-words in the process of sequence generation. Interruption leads to a sequence remaining in the form of subfragments. Thus, the unambiguous segments between branching points whose order is not uniquely determined are called sequence subfragments.

Informative fragments are defined as fragments of a sequence that are determined by the nearest ends of overlapped physical sequence fragments.

A certain number of physical fragments may be pooled without losing the possibility of defining informative fragments. The total length of randomly pooled fragments depends on the length of k-tuples that are used in the sequencing process.

The algorithm consists of two main units. The first part is used for generation of subfragments from the set of k-tuples contained in a sequence. Subfragments may be generated within the coding region of physical nucleic acid sequence of certain sizes, or within the informative fragments defined within long nucleic acid sequences. Both types of fragments are members of the basic library. This algorithm does not describe the determination of the content of the k-tuples of the informative fragments of the basic library, i.e. the step of preparation of informative fragments to be used in the sequence generation process.

The second part of the algorithm determines the linear order of obtained subfragments with the purpose of regenerating the complete sequence of the nucleic acid fragments of the basic library. For this purpose a second, ordering library is used, made of randomly pooled fragments of the starting sequence. The algorithm does not include the step of combining sequences of basic fragments to regenerate an entire, megabase plus sequence. This may be accomplished using the link-up of fragments of the basic library which is a prerequisite for informative fragment generation. Alternatively, it may be accomplished after generation of sequences of fragments of the basic library by this

algorithm, using search for their overlap, based on the presence of common end-sequences.

The algorithm requires neither knowledge of the number of appearances of a given k-tuple in a nucleic acid sequence of the basic and ordering libraries, nor does it require the information of which k-tuple words are present on the ends of a fragment. The algorithm operates with the mixed content of k-tuples of various length. The concept of the algorithm enables operations with the k-tuple sets that contain false positive and false negative k-tuples. Only in specific cases does the content of the false k-tuples primarily influence the completeness and correctness of the generated sequence. The algorithm may be used for optimization of parameters in simulation experiments, as well as for sequence generation in the actual SBH experiments e.g., generation of the genomic DNA sequence. In optimization of parameters, the choice of the oligonucleotide probes (k-tuples) for practical and convenient fragments and/or the choice of the optimal lengths and the number of fragments for the defined probes are especially important.

This part of the algorithm has a central role in the process of the generation of the e from the content of k-tuples. It is based on the unique ordering of k-tuples by means of maximal overlap. The main obstacles in sequence generation are specific repeated sequences and false positive and/or negative k-tuples. The aim of this part of the algorithm is to obtain the minimal number of the longest possible subfragments, with correct sequence. This part of the algorithm consists of one basic, and several control steps. A two-stage process is necessary since certain information can be used only after generation of all primary subfragments.

The main problem of sequence generation is obtaining a repeated sequence from word contents that by definition do not carry information on the number of occurrences of the particular k-tuples. The concept of the entire algorithm depends on the basis on which this problem is solved. In principle, there are two opposite approaches: 1) repeated sequences may be obtained at the beginning, in the process of generation of pSFs, or 2) repeated sequences can be obtained later, in the process of the final ordering of the subfragments. In the first case, pSFs contain an excess of sequences and in the second case, they contain a deficit of sequences. The first approach requires elimination

of the excess sequences generated, and the second requires permitting multiple use of some of the subfragments in the process of the final assembling of the sequence.

The difference in the two approaches in the degree of strictness of the rule of unique overlap of k-tuples. The less severe rule is: k-tuple X is unambiguously maximally overlapped with k-tuple Y if and only if, the rightmost k-1 end of k-tuple X is present only on the leftmost end of k-tuple Y. This rule allows the generation of repetitive sequences and the formation of surplus sequences.

A stricter rule which is used in the second approach has an addition caveat: k-tuple X is unambiguously maximally overlapped with k-tuple Y if and only if, the rightmost K-1 end of k-tuple X is present only on the leftmost end of k-tuple Y and if the leftmost K-l end of k-tuple Y is not present on the rightmost end of any other k-tuple. The algorithm based on the stricter rule is simpler, and is described herein.

The process of elongation of a given subfragment is stopped when the right k-1 end of the last k-tuple included is not present on the left end of any k-tuple or is present on two or more k-tuples. If it is present on only one k-tuple the second part of the rule is tested. If in addition there is a k-tuple which differs from the previously included one, the assembly of the given subfragment is terminated only on the first leftmost position. If this additional k-tuple does not exist, the conditions are met for unique k-1 overlap and a given subfragment is extended to the right by one element.

Beside the basic rule, a supplementary one is used to allow the usage of k-tuples of different lengths. The maximal overlap is the length of k-1 of the shorter k-tuple of the overlapping pair. Generation of the pSFs is performed starting from the first k-tuple from the file in which k-tuples are displayed randomly and independently from their order in a nucleic acid sequence. Thus, the first k-tuple in the file is not necessarily on the beginning of the sequence, nor on the start of the particular subfragment. The process of subfragment generation is performed by ordering the k-tuples by means of unique overlap, which is defined by the described rule. Each used k-tuple is erased from the file. At the point when there are no further k-tuples unambiguously overlapping with the last one included, the building of subfragment is terminated and the buildup of another pSF is started. Since generation of a majority of subfragments does not begin from their actual starts, the formed pSF are added to the k-tuple file and are considered as a longer k-tuple.

Another possibility is to form subfragments going in both directions from the starting k-tuple. The process ends when further overlap, i.e. the extension of any of the subfragments, is not possible.

The pSFs can be divided in three groups: 1) subfragments of the maximal length and correct sequence in cases of exact k-tuple set; 2) short subfragments, formed due to the used of the maximal and unambiguous overlap rule on the incomplete set, and/or the set with some false positive k-tuples; and 3) pSFs of an incorrect sequence. The incompleteness of the set in 2) is caused by false negative results of a hybridization experiment, as well as by using an incorrect set of k-tuples. These are formed due to the false positive and false negative k-tuples and can be: a) misconnected subfragments; b) subfragments with the wrong end; and c) false positive k-tuples which appears as false minimal subfragments.

Considering false positive k-tuples, there is the possibility for the presence of a k-tuple containing more than one wrong base or containing one wrong base somewhere in the middle, as well as the possibility for a k-tuple with a wrong base on the end. Generation of short, erroneous or misconnected subfragments is caused by the latter k-tuples. The k-tuples of the former two kinds represent wrong pSFs with length equal to k-tuple length.

In the case of one false negative k-tuple, pSFs are generated because of the impossibility of maximal overlapping. In the case of the presence of one false positive k-tuple with the wrong base on its leftmost or rightmost end, pSFs are generated because of the impossibility of unambiguous overlapping. When both false positive and false negative k-tuples with a common k-1 sequence are present in the file, pSFs are generated, and one of these pSFs contains the wrong k-tuple at the relevant end.

The process of correcting subfragments with errors in sequence and the linking of unambiguously connected pSF is performed after subfragment generation and in the process of subfragment ordering. The first step which consists of cutting the misconnected pSFs and obtaining the final subfragments by unambiguous connection of pSFs is described below.

There are two approaches for the formation of misconnected subfragments. In the first a mistake occurs when an erroneous k-tuple appears on the points of assembly of the

repeated sequences of lengths k-1. In the second, the repeated sequences are shorter than k-l. These situations can occur in two variants each. In the first variant, one of the repeated sequences represents the end of a fragment. In the second variant, the repeated sequence occurs at any position within the fragment. For the first possibility, the absence of some k-tuples from the file (false negatives) is required to generate a misconnection. The second possibility requires the presence of both false negative and false positive k-tuples in the file. Considering the repetitions of k-1 sequence, the lack of only one k-tuple is sufficient when either end is repeated internally. The lack of two is needed for strictly internal repetition. The reason is that the end of a sequence can be considered informatically as an endless linear array of false negative k-tuples. From the "smaller than k-l case", only the repeated sequence of the length of k-2, which requires two or three specific erroneous k-tuples, will be considered. It is very likely that these will be the only cases which will be detected in a real experiment, the others being much less frequent.

Recognition of the misconnected subfragments is more strictly defined when a repeated sequence does not appear at the end of the fragment. In this situation, one can detect further two subfragments, one of which contains on its leftmost, and the other on its rightmost end k-2 sequences which are also present in the misconnected subfragment. When the repeated sequence is on the end of the fragment, there is only one subfragment which contains k-2 sequence causing the mistake in subfragment formation on its leftmost or rightmost end.

The removal of misconnected subfragments by their cutting is performed according to the common rule: If the leftmost or rightmost sequence of the length of k-2 of subfragments is present in any other subfragment, the subfragment is to be cut into subfragments, each of them containing k-2 sequence. This rule does not cover rarer situations of a repeated end when there are more than one false negative k-tuple on the point of repeated k-1 sequence. Misconnected subfragments of this kind can be recognized by using the information from the overlapped fragments, or informative fragments of basic and ordering libraries. In addition, the misconnected subfragment will remain when two or more false negative k-tuples occur on both positions which contain the identical k-1 sequence. This is a very rare situation since it requires at least 4 specific

false k-tuples. An additional rule can be introduced to cut these subfragments on sequences of length k if the given sequence can be obtained by combination of sequences shorter than k-2 from the end of one subfragment and the start of another.

By strict application of the described rule, some completeness is lost to ensure the accuracy of the output. Some of the subfragments will be cut although they are not misconnected since they fit into the pattern of a misconnected subfragment. There are several situations of this kind. For example, a fragment, beside at least two identical k-1 sequences, contains any k-2 sequence from k-1 or a fragment contains k-2 sequence repeated at least twice and at least one false negative k-tuple containing given k-2 sequence in the middle, etc.

The aim of this part of the algorithm is to reduce the number of pSFs to a minimal number of longer subfragments with correct sequence. The generation of unique longer subfragments or a complete sequence is possible in two situations. The first situation concerns the specific order of repeated k-1 words. There are cases in which some or all maximally extended pSFs (the first group of pSFs) can be uniquely ordered. For example, in fragment S-Rl-a-R2-b-R1-c-R2-E where S and E are the start and end of a fragment, a, b, and c are different sequences specific to respective subfragments and R1 and R2 are two k-1 sequences that are tandemly repeated, five subfragments are generated (S-R1, R1-a-R2, R2-b-R1, Rl-c-R2, and R-E). They may be ordered in two ways; the original sequence above or S-R1-c-R-b-R1-a-R-E. In contrast, in a fragment with the same number and types of repeated sequences but ordered differently, i.e. S-R1-a-R1-b-R-c-R-E, there is no other sequence which includes all subfragments. Examples of this type can be recognized only after the process of generation of pSFs. They represent the necessity for two steps in the process of pSF generation. The second situation of generation of false short subfragments on positions of nonrepeated k-1 sequences when the files contain false negative and/or positive k-tuples is more important.

The solution for both pSF groups consists of two parts. First, the false positive k-tuples appearing as the nonexisting minimal subfragments are eliminated. All k-tuple subfragments of length k which do not have an overlap on either end, of the length of longer than k-a on one end and longer than k-b on the other end, are eliminated to enable

formation of the maximal number of connections. In our experiments, the values for a and b of 2 and 3, respectively, appeared to be adequate to eliminate a sufficient number of false positive k-tuples.

The merging of subfragments that can be uniquely connected is accomplished in the second step. The rule for connection is: two subfragments may be unambiguously connected if, and only if, the overlapping sequence at the relevant end or start of two subfragments is not present at the start and/or end of any other subfragment.

The exception is if one subfragment from the considered pair has the identical beginning and end. In that case connection is permitted, even if there is another subfragment with the same end present in the file. The main problem here is the precise definition of overlapping sequence. The connection is not permitted if the overlapping sequence unique for only one pair of subfragments is shorter than k-2, or it is k-2 or longer but an additional subfragment exists with the overlapping sequence of any length longer than k-4. Also, both the canonical ends of pSFs and the ends after omitting one (or few) last bases are considered as the overlapping sequences.

After this step some false positive k-tuples (as minimal subfragments) and some subfragments with a wrong end may survive. In addition, in very rare occasions where a certain number of some specific false k-tuples are simultaneously present, an erroneous connection may take place. These cases will be detected and solved in the subfragment ordering process, and in the additional control steps along with the handling of uncut "misconnected" subfragments.

The short subfragments that are obtained are of two kinds. In the common case, these subfragments may be unambiguously connected among themselves because of the distribution of repeated k- 1 sequences. This may be done after the process of generation of pSFs and is a good example of the necessity for two steps in the process of pSF generation. In the case of using the file containing false positive and/or false negative k-tuples, short pSFs are obtained on the sites of nonrepeated k-1 sequences. Considering false positive k-tuples, a k-tuple may contain more than one wrong base (or containing one wrong base somewhere in the middle), as well as k-tuple on the end. Generation of short and erroneous (or misconnected) subfragments is caused by the latter k-tuples. The k-tuples of the former kind represent wrong pSFs with length equal to k- tuple length.

The aim of merging pSF part of the algorithm is the reduction of the number of pSFs to the minimal number of longer subfragments with the correct sequence. All k-tuple subfragments that do not have an overlap on either end, of the length of longer than k-a on one, and longer than k-b on the other end, are eliminated to enable the maximal number of connections. In this way, the majority of false positive k-tuples are discarded. The rule for connection is: two subfragments can be unambiguously connected if, and only if the overlapping sequence of the relevant end or start of two subfragments is not present on the start and/or end of any other subfragment. The exception is a subfragment with the identical beginning and end. In that case connection is permitted, provided that there is another subfragment with the same end present in the file. The main problem here is of precise definition of overlapping sequence. The presence of at least two specific false negative k-tuples on the points of repetition of k-1 or k-2 sequences, as well as combining of the false positive and false negative k-tuples may destroy or "mask" some overlapping sequences and can produce an unambiguous, but wrong connection of pSFs. To prevent this, completeness must be sacrificed on account of exactness: the connection is not permitted on the end-sequences shorter than k-2, and in the presence of an extra overlapping sequence longer than k-4. The overlapping sequences are defined from the end of the pSFs, or omitting one, or few last bases.

In the very rare situations, with the presence of a certain number of some specific false positive and false negative k-tuples, some subfragments with the wrong end can survive, some false positive k-tuples (as minimal subfragments) can remain, or the erroneous connection can take place. These cases are detected and solved in the subfragments ordering process, and in the additional control steps along with the handling of uncut, misconnected subfragments.

The process of ordering of subfragments is similar to the process of their generation. If one considers subfragments as longer k-tuples, ordering is performed by their unambiguous connection via overlapping ends. The informational basis for unambiguous connection is the division of subfragments generated in fragments of the basic library into groups representing segments of those fragments. The method is analogous to the biochemical solution of this problem based on hybridization with longer oligonucleotides with relevant connecting sequence. The connecting sequences are

generated as subfragments using the k-tuple sets of the appropriate segments of basic library fragments. Relevant segments are defined by the fragments of the ordering library that overlap with the respective fragments of the basic library. The shortest segments are informative fragments of the ordering library. The longer ones are several neighboring informative fragments or total overlapping portions of fragments corresponding of the ordering and basic libraries. In order to decrease the number of separate samples, fragments of the ordering library are randomly pooled, and the unique k-tuple content is determined.

By using the large number of fragments in the ordering library very short segments are generated, thus reducing the chance of the multiple appearance of the k-1 sequences which are the reasons for generation of the subfragments. Furthermore, longer segments, consisting of the various regions of the given fragment of the basic library, do not contain some of the repeated k-1 sequences. In every segment a connecting sequence (a connecting subfragment) is generated for a certain pair of the subfragments from the given fragment. The process of ordering consists of three steps: (I) generation of the k-tuple contents of each segment; (2) generation of subfragments in each segment; and (3) connection of the subfragments of the segments. Primary segments are defined as significant intersections and differences of k-tuple contents of a given fragment of the basic library with the k-tuple contents of the pools of the ordering library. Secondary (shorter) segments are defined as intersections and differences of the k-tuple contents of the primary segments.

There is a problem of accumulating both false positive and negative k-tuples in both the differences and intersections. The false negative k-tuples from starting sequences accumulate in the intersections (overlapping parts), as well as false positive k-tuples occurring randomly in both sequences, but not in the relevant overlapping region. On the other hand, the majority of false positives from either of the starting sequences is not taken up into intersections. This is an example of the reduction of experimental errors from individual fragments by using information from fragments overlapping with them. The false k-tuples accumulate in the differences for another reason. The set of false negatives from the original sequences are enlarged for false positives from intersections and the set of false positives for those k-tuples which are not included in the

intersection by error, i.e. are false negative in the intersection. If the starting sequences contain 10% false negative data, the primary and secondary intersections will contain 19% and 28% false negative k-tuples. respectively. On the other hand, a mathematical expectation of 77 false positives may be predicted if the basic fragment and the pools have lengths of 500 bp and 10,000 bp, respectively. However, there is a possibility of recovering most of the "lost" k-tuples and of eliminating most of the false positive k-tuples.

First, one has to determine a basic content of the k-tuples for a given segment as the intersection of a given pair of the k-tuple contents. This is followed by including all k-tuples of the starting k-tuple contents in the intersection, which contain at one end k-1 and at the other end k-+ sequences which occur at the ends of two k-tuples of the basic set. This is done before generation of the differences thus preventing the accumulation of false positives in that process. Following that, the same type of enlargement of k-tuple set is applied to differences with the distinction that the borrowing is from the intersections. All borrowed k-tuples are eliminated from the intersection files as false positives.

The intersection, i.e. a set of common k-tuples, is defined for each pair (a basic fragment) x (a pool of ordering library). If the number of k-tuples in the set is significant it is enlarged with the false negatives according to the described rule. The primary difference set is obtained by subtracting from a given basic fragment the obtained intersection set. The false negative k-tuples are appended to the difference set by borrowing from the intersection set according to the described rule and, at the same time, removed from the intersection set as false positive k-tuples. When the basic fragment is longer than the pooled fragments, this difference can represent the two separate segments which somewhat reduces its utility in further steps. The primary segments are all generated intersections and differences of pairs (a basic fragment) x (a pool of ordering library) containing the significant number of k-tuples. K-tuple sets of secondary segments are obtained by comparison of k-tuple sets of all possible pairs of primary segments. The two differences are defined from each pair which produces the intersection with the significant number of k-tuples. The majority of available

information from overlapped fragments is recovered in this step so that there is little to be gained from the third round of forming intersections, and differences.

(2) Generation of the subfragments of the segments is performed identically as described for the fragments of the basic library.

(3) The method of connection of subfragments consists of sequentially determining the correctly linked pairs of subfragments among the subfragments from a given basic library fragment which have some overlapped ends. In the case of 4 relevant subfragments, two of which contain the same beginning and two having the same end, there are 4 different pairs of subfragments that can be connected. In general 2 are correct and 2 are wrong. To find correct ones, the presence of the connecting sequences of each pair is tested in the subfragments generated from all primary and secondary segments for a given basic fragment. The length and the position of the connecting sequence are chosen to avoid interference with sequences which occur by chance. They are k+2 or longer, and include at least one element 2 beside overlapping sequence in both subfragments of a given pair. The connection is permitted only if the two connecting sequences are found and the remaining two do not exist. The two linked subfragments replace former subfragments in the file and the process is cyclically repeated. Repeated sequences are generated in this step. This means that some subfragments are included in linked subfragments more than once. They will be recognized by finding the relevant connecting sequence which engages one subfragment in connection with two different subfragments.

The recognition of misconnected subfragments generated in the processes of building pSFs and merging pSFs into longer subfragments is based on testing whether the sequences of subfragments from a given basic fragment exist in the sequences of subfragments generated in the segments for the fragment. The sequences from an incorrectly connected position will not be found indicating the misconnected subfragments.

Beside the described three steps in ordering of subfragments some additional control steps or steps applicable to specific sequences will be necessary for the generation of more complete sequence without mistakes.

The determination of which subfragment belongs to which segment is performed by comparison of contents of k-tuples in segments and subfragments. Because of the errors in the k-tuple contents (due to the primary error in pools and statistical errors due to the frequency of occurrences of k-tuples) the exact partitioning of subfragments is impossible. Thus, instead of "all or none" partition, the chance of coming from the given segment (P(sf,s)) is determined for each subfragment. This possibility is the function of the lengths of k-tuples, the lengths of subfragments, the lengths of fragments of ordering library, the size of the pool, and of the percentage of false k-tuples in the file:

$$P(sf,s)=(Ck-F)/Lsf,$$

where Lsf is the length of subfragment, Ck is the number of common k-tuples for a given subfragment/segment pair, and F is the parameter that includes relations between lengths of k-tuples, fragments of basic library, the size of the pool, and the error percentage.

Subfragments attributed to a particular segment are treated as redundant short pSFs and are submitted to a process of unambiguous connection. The definition of unambiguous connection is slightly different in this case, since it is based on a probability that subfragments with overlapping end(s) belong to the segment considered. Besides, the accuracy of unambiguous connection is controlled by following the connection of these subfragments in other segments. After the connection in different segments, all of the obtained subfragments are merged together, shorter subfragments included within longer ones are eliminated, and the remaining ones are submitted to the ordinary connecting process. If the sequence is not regenerated completely, the process of partition and connection of subfragments is repeated with the same or less severe criterions of probability of belonging to the particular segment, followed by unambiguous connection.

Using severe criteria for defining unambiguous overlap, some information is not used. Instead of a complete sequence, several subfragments that define a number of possibilities for a given fragment are obtained. Using less severe criteria an accurate and complete sequence is generated. In a certain number of situations, e.g., an erroneous connection, it is possible to generate a complete, but an incorrect sequence, or to generate "monster" subfragments with no connection among them . Thus, for each fragment of the basic library one obtains: a) several possible solutions where one is correct and b) the

most probable correct solution. Also, in a very small number of cases, due to the mistake in the subfragment generation process or due to the specific ratio of the probabilities of belonging. no unambiguous solution is generated or one, the most probable solution. These cases remain as incomplete sequences, or the unambiguous solution is obtained by comparing these data with other, overlapped fragments of basic library.

The described algorithm was tested on a randomly generated, 50 kb sequence, containing 40% GC to simulate the GC content of the human genome. In the middle part of this sequence were inserted various All, and some other repetitive sequences, of a total length of about 4 kb. To simulate an *in vitro* SBH experiment, the following operations were performed to prepare appropriate data.

1) Positions of sixty 5 kb overlapping "clones" were randomly defined, to simulate preparation of a basic library:

2) Positions of one thousand 500 bp "clones" were randomly determined to simulate making the ordering library. These fragments were extracted from the sequence. Random pools of 20 fragments were made, and k-tuple sets of pools were determined and stored on the hard disk. These data are used in the subfragment ordering phase: For the same density of clones 4 million clones in basic library and 3 million clones in ordering library are used for the entire human genome. The total number of 7 million clones is several fold smaller than the number of clones a few kb long for random cloning of almost all of genomic DNA and sequencing by a gel-based method.

From the data on the starts and ends of 5 kb fragments, 117 "informative fragments" were determined to be in the sequence. This was followed by determination of sets of overlapping k-tuples of which the single "informative fragment" consist. Only the subset of k-tuples matching a predetermined list were used. The list contained 65% 8-mers, 30% 9-mers, and 5% 10-1 2-mers. Processes of generation and the ordering of subfragments were performed on these data.

The testing of the algorithm was performed on the simulated data in two experiments. The sequence of 50 informative fragments was regenerated with the 100% correct data set (over 20,000 bp), and 26 informative fragments (about 10,000 bp) with 10% false k-tuples (5% positive and 5% negative ones).

In the first experiment, all subfragments were correct and in only one out of 50 informative fragments the sequence was not completely regenerated but remained in the form of 5 subfragments. The analysis of positions of overlapped fragments of ordering library has shown that they lack the information for the unique ordering of the 5 subfragments. The subfragments may be connected in two ways based on overlapping ends, 1-2-3-4-5 and 1-4-3-2-5. The only difference is the exchange of positions of subfragments 2 and 4. Since subfragments 2, 3, and 4 are relatively short (total of about 100 bp), the relatively greater chance existed, and occurred in this case, that none of the fragments of ordering library started or ended in the subfragment 3 region.

To simulate real sequencing, some false ("hybridization") data was included as input in a number of experiments. In oligomer hybridization experiments, under proposed conditions, the only situation producing unreliable data is the end mismatch versus full match hybridization. Therefore, in simulation only those k-tuples differing in a single element on either end from the real one were considered to be false positives. These "false" sets are made as follows. On the original set of a k-tuples of the informative fragment, a subset of 5% false positive k-tuples are added. False positive k-tuples are made by randomly picking a k-tuple from the set, copying it and altering a nucleotide on its beginning or end. This is followed by subtraction of a subset of 5% randomly chosen k-tuples. In this way the statistically expected number of the most complicated cases is generated in which the correct k-tuple is replaced with a k-tuple with the wrong base on the end.

Production of k-tuple sets as described leads to up to 10% of false data. This value varies from case to case, due to the randomness of choice of k-tuples to be copied, altered, and erased. Nevertheless, this percentage 3-4 times exceeds the amount of unreliable data in real hybridization experiments. The introduced error of 10% leads to the two fold increase in the number of subfragments both in fragments of basic library (basic library informative fragments) and in segments. About 10% of the final subfragments have a wrong base at the end as expected for the k-tuple set which contains false positives (see generation of primary subfragments). Neither the cases of misconnection of subfragments nor subfragments with the wrong sequence were observed. In 4 informative fragments out of 26 examined in the ordering process the

complete sequence was not regenerated. In all 4 cases the sequence was obtained in the form of several longer subfragments and several shorter subfragments contained in the same segment. This result shows that the algorithmic principles allow working with a large percentage of false data.

The success of the generation of the sequence from its k-tuple content may be described in terms of completeness and accuracy. In the process of generation, two particular situations can be defined: 1) Some part of the information is missing in the generated sequence, but one knows where the ambiguities are and to which type they belong, and 2) the regenerated sequence that is obtained does not match the sequence from which the k- tuple content is generated, but the mistake can not be detected. Assuming the algorithm is developed to its theoretical limits, as in the use of the exact k-tuple sets, only the first situation can take place. There the incompleteness results in a certain number of subfragments that may not be ordered unambiguously and the problem of determination of the exact length of monotonous sequences, i.e. the number of perfect tandem repeats.

With false k-tuples, incorrect sequences may be generated. The reason for mistakes does not lie in the shortcomings of the algorithm, but in the fact that a given content of k-tuples unambiguously represents the sequence that differs from the original one. One may define three classes of error, depending on the kind of the false k- tuples present in the file. False negative k-tuples (which are not accompanied with the false. positives) produce "deletions". False positive k-tuples are producing "elongations (unequal crossing over)". False positives accompanied with false negatives are the reason for generation of "insertions", alone or combined with "deletions". The deletions are produced when all of the k-tuples (or their majority) between two possible starts of the subfragments are false negatives. Since every position in the sequence is defined by k-tuples, the occurrence of the deletions in a common case requires k consecutive false negatives. (With 10% of the false negatives and k=8, this situation takes place after every 108 elements). This situation is extremely infrequent even in mammalian genome sequencing using random libraries containing ten genome equivalents.

Elongation of the end of the sequence caused by false positive k-tuples is the special case of "insertions", since the end of the sequence can be considered as the

endless linear array of false negative k-tuples. One may consider a group of false positive k-tuples producing subfragments longer than one k-tuple. Situations of this kind may be detected if subfragments are generated in overlapped fragments, like random physical fragments of the ordering library. An insertion, or insertion in place of a deletion, can arise as a result of specific combinations of false positive and false negative k-tuples. In the first case. the number of consecutive false negatives is smaller than k. Both cases require several overlapping false positive k-tuples. The insertions and deletions are mostly theoretical possibilities without sizable practical repercussions since the requirements in the number and specificity of false k-tuples are simply too high.

In every other situation of not meeting the theoretical requirement of the minimal number an the kind of the false positive and/or negatives, mistakes in the k-tuples content may produce only the lesser completeness of a generated sequence.

In SBH, a sample nucleic acid is sequenced by exposing the sample to a support-bound probe of known sequence and a labeled probe or probes in solution. Wherever the probes ligase is introduced into the mixture of probes and sample, such that, wherever a support has a bound probe and a labeled probe hybridized back to back along the sample, the two probes will be chemically linked by the action of the ligase. After washing, only chemically linked support-bound and labeled probes are detected by the presence of the labeled probe. By knowing the identity of the support-bound probe at a particular location in an array, and the identity of the labeled probe, a portion of the sequence of the sample may be determined by the presence of a label at a point in an array on a Format with a sample of three substrates Not all of the sample to be sequenced may be a nucleic acid fragment or oligonucleotide of ten base pairs ("bp"). The sample is preferably four to one thousand bases in length.

The length of the probe is a fragment less than ten bases in length, and, preferably, is between four and nine bases in length. In this way, arrays of support-bound probes may include all oligonucleotides of a given length or may include only oligonucleotides selected for a particular test. Where all oligonucleotides of a given length are used, the number of central oligonucleotides may be calculated by 4^N where N is the length of the probe.

5.36 RE-USING SEQUENCING CHIPS

When ligation is employed in the sequencing process, then the ordinary oligonucleotide chip cannot be immediately reused. The inventors contemplate that this may be overcome in various ways.

One may employ ribonucleotides for the second probe, probe P, so that this probe may subsequently be removed by RNAse treatment. RNAse treatment may utilize RNAse A, an endoribonuclease that specifically attacks single-stranded RNA 3'-to pyrimidine residues and cleaves the phosphate linkage to the adjacent nucleotide. The end products are pyrimidine 3' phosphates and oligonucleotides with terminal pyrimidine 3' phosphates. RNAse A works in the absence of cofactors and divalent cations.

To utilize an RNAse, one would generally incubate the chip in any appropriate RNAse-containing buffer, as described by Sambrook *et al.*. (1989; incorporated herein by reference). The use of 30-50 µl of RNAse-containing buffer per 8 x 8 mm or 9 x 9mm array at 37°C for between 10 and 60 minutes is appropriate. One would then wash with hybridization buffer.

Although not widely applicable, one could also use the uracil base, as described by Craig *et al.*. (1989), incorporated herein by reference, in specific embodiments. Destruction of the ligated probe combination, to yield a re-usable chip, would be achieved by digestion with *the E. coli* repair enzyme, uracil-DNA glycosylase which removes uracil from DNA.

One could also generate a specifically cleavable bond between the probes and then cleave the bond after detection. For example, this may be achieved by chemical ligation as described by Shabarova *et al.*., (1991) and Dolmnnaya et at, (1988), both references being specifically incorporated herein by reference.

Shabarova *et al.*. (1991) describe the condensation of oligodeoxyribonucleotides with cyanogen bromide as a condensing agent. In their one step chemical ligation reaction, the oligonucleotides are heated to 97°C, slowly cooled to 0°C, then 1 µl 10mM BrCN in acetonitrile is added.

Dolmnnaya *et al.*. (1988) show how to incorporate phosphoramidite and pyrophosphate internucleotide bonds in DNA duplexes. They also use a chemical ligation method for modification of the sugar phosphate backbone of DNA, with a water-soluble

carbodiimide (CDI) as a coupling agent. The selective cleavage of a phosphoamide bond involves contact with 15% CH_3COOH for 5 mm at 95°C. The selective cleavage of a pyrophosphate bond involves contact with a pyridine-water mixture (9:1) and freshly distilled $(CF_3CO)_2O$.

5.37 DIAGNOSTICS - SCORING KNOWN MUTATIONS OR FULL GENE RESEQUENCING

In a simple case, the goal may be to discover whether selected, known mutations occur in a DNA segment. Less than 12 probes may suffice for this purpose, for example, 5 probes positive for one allele, 5 positive for the other, and 2 negative for both. Because of the small number of probes to be scored per sample, large numbers of samples may be analyzed in parallel. For example, with 12 probes in 3 hybridization cycles, 96 different genomic loci or gene segments from 64 patient may be analyzed on one 6 x 9 in membrane containing 12 x 24 subarrays each with 64 dots representing the same DNA segment from 64 patients. In this example, samples may be prepared in sixty-four 96-well plates. Each plate may represent one patient, and each well may represent one of the DNA segments to be analyzed. The samples from 64 plates may be spotted in four replicas as four quarters of the same membrane.

A set of 12 probes may be selected by single channel pipetting or by a single pin transferring device (or by an array of individually-controlled pipets or pins) for each of the 96 segments, and the selected probes may be arrayed in twelve 96-well plates. Probes may be labelled, if they are not prelabelled, and then probes from four plates may be mixed with hybridization buffer and added to the subarrays preferentially by a 96-channel pipeting device. After one hybridization cycle it is possible to strip off previously-applied probes by incubating the membrane at 37° to 55°C in the preferably undiluted hybridization or washing buffer.

The likelihood that probes positive for one allele are positive and probes positive for the other allele are negative may be used to determine which of the two alleles is present. In this redundant scoring scheme, some level (about 10%) of errors in hybridization of each probe may be tolerated.

An incomplete set of probes may be used for scoring most of the alleles, especially if the smaller redundancy is sufficient, e.g., one or two probes which prove the presence or absence in a sample of one of the two alleles. For example, with a set of four thousand 8-mers there is a 91% chance of finding at least one positive probe for one of the two alleles for a randomly selected locus. The incomplete set of probes may be optimized to reflect G+C content and other biases in the analyzed samples.

For full gene sequencing, genes may be amplified in an appropriate number of segments. For each segment, a set of probes (about one probe per 2-4 bases) may be selected and hybridized. These probes may identify whether there is a mutation anywhere in the analyzed segments. Segments (*i.e.,* subarrays which contain these segments) where one or more mutated sites are detected may be hybridized with additional probes to find the exact sequence at the mutated sites. If a DNA sample is tested by every second 6-mer, and a mutation is localized at the position that is surrounded by positively hybridized probes TGCAAA and TATTCC and covered by three negative probes: CAAAAC, AAACTA and ACTATT, the mutated nucleotides must be A and/or C occurring in the normal sequence at that position. They may be changed by a single base mutation, or by a one or two nucleotide deletion and/or insertion between bases AA, AC or CT.

One approach is to select a probe that extends the positively hybridized probe TGCAAA for one nucleotide to the right, and which extends the probe TATTCC one nucleotide to the left, With these 8 probes (GCAAAA, GCAAAT, GCAAAC, GCAAAG and ATATTC, TTATTC, CTATTC, GTATTC) two questionable nucleotides are determined.

The most likely hypothesis about the mutation may be determined. For example, A is found to be mutated to G. There are two solutions satisfied by these results. Either replacement of A with G is the only change or there is in addition to that change an insertion of some number of bases between newly determined G and the following C. If the result with bridging probes is negative these options may then be checked first by at least one bridging probe comprising the mutated position (AAGCTA) and with an additional 8 probes: CAAAGA, CAAAGT, CAAAGC, CAAAGG and ACTATT, TCTATT, CCTATT, GCTATT, There are many other ways to select mutation-solving probes.

In the case of diploid, particular comparisons of scores for the test samples and homozygotic control may be performed to identify heterozygotes (see above). A few consecutive probes are expected to have roughly twice smaller signals if the segment covered by these probes is mutated on one of the two chromosomes.

5.38 IDENTIFICATION OF GENES (MUTATIONS) RESPONSIBLE FOR GENETICDISORDERS AND OTHER TRAITS

Using universal sets of longer probes (8-mers or 9-mers) on immobilized arrays of samples, DNA fragments as long as 5-20 kb may be sequenced without subcloning. Furthermore, the speed of sequencing readily may be about 10 million bp/day/hybridization instrument. This performance allows for resequencing a large fraction of human genes or the human genome repeatedly from scientifically or medically interesting individuals. To resequence 50% of the human genes, about 100 million bp is checked. That may be done in a relatively short period of time at an affordable cost.

This enormous resequencing capability may be used in several ways to identify mutations and/or genes that encode for disorders or any other traits. Basically, mRNAs (which may be converted into cDNAs) from particular tissues or genomic DNA of patients with particular disorders may be used as starting materials. From both sources of DNA, separate genes or genomic fragments of appropriate length may be prepared either by cloning procedures or by *in vitro* amplification procedures (for example by PCR). If cloning is used, the minimal set of clones to be analyzed may be selected from the libraries before sequencing. That may be done efficiently by hybridization of a small number of probes, especially if a small number of clones longer than 5 kb is to be sorted. Cloning may increase the amount of hybridization data about two times, but does not require tens of thousands of PCR primers.

In one variant of the procedure, gene or genomic fragments may be prepared by restriction cutting with enzymes like Hga I which cuts DNA in following way: GACGC(N5')/CTGCG(N10'). Protruding ends of five bases are different for different fragments. One enzyme produces appropriate fragments for a certain number of genes. By cutting cDNA or genomic DNA with several enzymes in separate reactions, every gene of interest may be excised appropriately. In one approach, the cut DNA is

fractionated by size. DNA fragments prepared in this way (and optionally treated with Exonuclease III which individually removes nucleotides from the 3' end and increases length and specificity of the ends) may be dispensed in the tubes or in multiwell plates. From a relatively small set of DNA adapters with a common portion and a variable protruding end of appropriate length, a pair of adapters may be selected for every gene fragment that needs to be amplified. These adapters are ligated and then PCR is performed by universal primers. From 1000 adapters, a million pairs may be generated, thus a million different fragments may be specifically amplified in the identical conditions with a universal pair of primers complementary to the common end of the adapters.

If a DNA difference is found to be repeated in several patients, and that sequence change is nonsense or can change function of the corresponding protein, then the mutated gene may be responsible for the disorder. By analyzing a significant number of individuals with particular traits, functional allelic variations of particular genes could be associated by specific traits.

This approach may be used to eliminate the need for very expensive genetic mapping on extensive pedigrees and has special value when there is no such genetic data or material.

5.39 SCORING SINGLE NUCLEOTIDE POLYMORPHISMS IN GENETIC MAPPING

Techniques disclosed in this application are appropriate for an efficient identification of genomic fragments with single nucleotide polymorphisms (SNUPs). In 10 individuals by applying the described sequencing process on a large number of genomic fragments of known sequence that may be amplified by cloning or by *in vitro* amplification, a sufficient number of DNA segments with SNUPs may be identified. The polymorphic fragments are further used as SNUP markers. These markers are either mapped previously (for example they represent mapped STSs) or they may be mapped through the screening procedure described below.

SNUPs may be scored in every individual from relevant families or populations by amplifying markers and arraying them in the form of the array of subarrays. Subarrays

contain the same marker amplified from the analyzed individuals. For each marker, as in the diagnostics of known mutations, a set of 6 or less probes positive for one allele and 6 or less probes positive for the other allele may be selected and scored. From the significant association of one or a group of the markers with the disorder, chromosomal position of the responsible gene(s) may be determined. Because of the high throughput and low cost, thousands of markers may be scored for thousands of individuals. This amount of data allows localization of a gene at a resolution level of less than one million bp as well as localization of genes involved in polygenic diseases. Localized genes may be identified by sequencing particular regions from relevant normal and affected individuals to score a mutation(s).

PCR is preferred for amplification of markers from genomic DNA. Each of the markers require a specific pair of primers. The existing markers may be convertible or new markers may be defined which may be prepared by cutting genomic DNA by Hga I type restriction enzymes, and by ligation with a pair of adapters.

SNUP markers can be amplified or spotted as pools to reduce the number of independent amplification reactions. In this case, more probes are scored per one sample. When 4 markers are pooled and spotted on 12 replica membranes, then 48 probes (12 per marker) may be scored in 4 cycles.

5.40 DETECTION AND VERIFICATION OF IDENTITY OF DNA FRAGMENTS

DNA fragments generated by restriction cutting, cloning or *in vitro* amplification (e.g. PCR) frequently may be identified in a experiment. Identification may be performed by verifying the presence of a DNA band of specific size on gel electrophoresis.

Alternatively, a specific oligonucleotide may be prepared and used to verify a DNA sample in question by hybridization. The procedure developed here allows for more efficient identification of a large number of samples without preparing a specific oligonucleotide for each fragment. A set of positive and negative probes may be selected from the universal set for each fragment on the basis of the known sequences. Probes that are selected to be positive usually are able to form one or a few overlapping groups and negative probes are spread over the whole insert.

This technology may be used for identification of STSs in the process of their mapping on the YAC clones. Each of the STSs may be tested on about 100 YAC clones or pools of YAC clones. DNAs from these 100 reactions possibly are spotted in one subarray. Different STSs may represent consecutive subarrays. In several hybridization cycles, a signature may be generated for each of the DNA samples, which signature proves or disproves existence of the particular STS in the given YAC clone with necessary confidence.

To reduce the number of independent PCR reactions or the number of independent samples for spotting, several STSs may be amplified simultaneously in a reaction or PCR samples may be mixed, respectively. In this case more probes have to be scored per one dot. The pooling of STSs is independent of pooling YACs and may be used on single YACs or pools of YACs. This scheme is especially attractive when several probes labelled with different colors are hybridized together.

In addition to confirmation of the existence of a DNA fragment in a sample, the amount of DNA may be estimated using intensities of the hybridization of several separate probes or one or more pools of probes. By comparing obtained intensities with intensities for control samples having a known amount of DNA, the quantity of DNA in all spotted samples is determined simultaneously. Because only a few probes are necessary for identification of a DNA fragment, and there are N possible probes that may be used for DNA N bases long, this application does not require a large set of probes to be sufficient for identification of any DNA segment. From one thousand 8-mers, on average about 30 full matching probes may be selected for a 1000 bp fragment.

5.41 IDENTIFICATION OF INFECTIOUS DISEASE ORGANISMS AND THEIR VARIANTS

DNA-based tests for the detection of viral, bacterial, fungal and other parasitic organisms in patients are usually more reliable and less expensive than alternatives. The major advantage of DNA tests is to be able to identify specific strains and mutants, and eventually be able to apply more effective treatment. Two applications are described below.

The presence of 12 known antibiotic resistance genes in bacterial infections may be tested by amplifying these genes. The amplified products from 128 patients may be spotted in two subarrays and 24 subarrays for 12 genes may then be repeated four times on a 8 x 12 cm membrane. For each gene, 12 probes may be selected for positive and negative scoring. Hybridizations may be performed in 3 cycles. For these tests, a much smaller set of probes is most likely to be universal. For example, from a set of one thousand 8-mers, on average 30 probes are positive in 1000 bp fragments, and 10 positive probes are usually sufficient for a highly reliable identification. As described in Section 5.22, several genes may be amplified and/or spotted together and the amount of the given DNA may be determined. The amount of amplified gene may be used as an indicator of the level of infection.

Another example involves possible sequencing of one gene or the whole genome of an HIV virus. Because of rapid diversification, the virus poses many difficulties for selection of an optimal therapy. DNA fragments may be amplified from isolated viruses from up to 64 patients and resequenced by the described procedure. On the basis of the obtained sequence the optimal therapy may be selected. If there is a mixture of two virus types of which one has the basic sequence (similar to the case of heterozygotes), the mutant may be identified by quantitative comparisons of its hybridization scores with scores of other samples, especially control samples containing the basic virus type only. Scores twice as small may be obtained for three to four probes that cover the site mutated in one of the two virus types present in the sample (see above).

5.42 FORENSIC AND PARENTAL IDENTIFICATION

Sequence polymorphisms make an individual genomic DNA unique. This permits analysis of blood or other body fluids or tissues from a crime scene and comparison with samples from criminal suspects. A sufficient number of polymorphic sites are scored to produce a unique signature of a sample. SBH may easily score single nucleotide polymorphisms to produce such signatures.

A set of DNA fragments (10-1000 mer) may be amplified from samples and suspects. DNAs from samples and suspects representing one fragment are spotted in one or several subarrays and each subarray may be replicated 4 times. In three cycles, 12

probes may determine the presence of allele A or B in each of the samples, including suspects, for each DNA locus. Matching the patterns of samples and suspects may lead to discovery of the suspect responsible for the crime.

The same procedure may be applicable to prove or disprove the identity of parents of a child. DNA may be prepared and polymorphic loci amplified from the child and adults; patterns of A or B alleles may be determined by hybridization for each. Comparisons of the obtained patterns, along with positive and negative controls, aide in the determination of familial relationships. In this case, only a significant portion of the alleles need match with one parent for identification. Large numbers of scored loci allow for the avoidance of statistical errors in the procedure or of masking effects of *de novo* mutations.

5.43 ASSESSING GENETIC DIVERSITY OF POPULATIONS OR SPECIES AND BIOLOGICAL DIVERSITY OF ECOLOGICAL NICHES

Measuring the frequency of allelic variations on a significant umber of loci (for example, several genes or entire mitochondrial DNA) permits development of different types of conclusions, such as conclusions regarding the impact of the environment on the genotypes, history and evolution of a population or its susceptibility to diseases or extinction, and others. These assessments may be performed by testing specific known alleles or by full resequencing of some loci to be able to define *de novo* mutations which may reveal fine variations or presence of mutagens in the environment.

Additionally, biodiversity in the microbial world may be surveyed by resequencing evolutionarily conserved DNA sequences, such as the genes for ribosomal RNAs or genes for highly conservative proteins. DNA may be prepared from the environment and particular genes amplified using primers corresponding to conservative sequences. DNA fragments may be cloned preferentially in a plasmid vector (or diluted to the level of one molecule per well in multiwell plates and than amplified *in vitro*). Clones prepared this way may be resequenced as described above. Two types of information are obtained. First of all, a catalogue of different species may be defined as well as the density of the individuals for each species. Another segment of information may be used to measure the influence of ecological factors or pollution on the ecosystem.

It may reveal whether some species are eradicated or whether the abundance ratios among species is altered due to the pollution. The method also is applicable for sequencing DNA from fossils.

5.44 DETECTION OR QUANTIFICATION OF NUCLEIC ACID SPECIES

DNA or RNA species may be detected and quantified by employing a probe pair including an unlabeled probe fixed to a substrate and a labeled probe in a solution. The species may be detected and quantified by exposure to the unlabeled probe in the presence of the labeled probe and ligase. Specifically, the formation of an extended probe by ligation of the labeled and unlabeled probe on the sample nucleic acid backbone is indicative of the presence of the species to be detected. Thus, the presence of label at a specific point in the array on the substrate after removing unligated labeled probe indicates the presence of a sample species while the quantity of label indicates the expression level of the species.

Alternatively, one or more unlabeled probes may be arrayed on a substrate as first members of pairs with one or more labeled probes to be introduced in solution. According to one method, multiplexing of the label on the array may be carried out by using dyes which fluoresce at distinguishable wavelengths. In this manner, a mixture of cDNAs applied to an array with pairs of labeled and unlabeled probes specific for species to be identified may be examined for the presence of and expression level of cDNA species. According to a preferred embodiment this approach may be carried out to sequence portions of cDNAs by selecting pairs of unlabeled and labeled probes pairs comprising sequences which overlap along the sequence of a cDNA to be detected.

Probes may be selected to detect the presence and quantity of particular pathogenic organisms genome by including in the composition selected probe pairs which appear in combination only in target pathogenic genome organisms. Thus, while no single probe pair may necessarily be specific for the pathogenic organism genome, the combination of pairs is. Similarly, in detecting or sequencing cDNAs, it might occur that a particular probe is not be specific for a cDNA or other type of species. Nevertheless, the presence and quantity of a particular species may be determined by a result wherein a

combination of selected probes situated at distinct array locations is indicative of the presence of a particular species.

An infectious agent with about 10kb or more of DNA may be detected using a support-bound detection chip without the use of polymerase chain reaction (PCR) or other target amplification procedures. According to other methods, the genomes of infectious agents including bacteria and viruses are assayed by amplification of a single target nucleotide sequence through PCR and detection of the presence of target by hybridization of a labelled probe specific for the target sequence. Because such an assay is specific for only a single target sequence it therefore is necessary to amplify the gene by methods such as PCR to provide sufficient target to provide a detectable signal.

According to this example, an improved method of detecting nucleotide sequences characteristic of infectious agents through a Format 3-type reaction is provided wherein a solid phase detection chip is prepared which comprises an array of multiple different immobilized oligonucleotide probes specific for the infectious agent of interest: A single dot comprising a mixture of many unlabeled probes complementary to the target nucleic acid concentrates the label specific to a species at one location thereby improving sensitivity over diffuse or single probe labeling. Such multiple probes may be of overlapping sequences of the target nucleotide sequence but may also be non-overlapping sequences as well as non-adjacent. Such probes preferably have a length of about 5 to 12 nucleotides.

A nucleic acid sample exposed to the probe array and target sequences present in the sample will hybridize with the multiple immobilized probes. A pool of multiple labeled probes selected to specifically bind to the target sequences adjacent to the immobilized probes is then applied with the sample to an array of unlabeled oligonucleotide probe mixtures. Ligase enzyme is then applied to the chip to ligate the adjacent probes on the sample. The detection chip is then washed to remove unhybridized and unligated probe and sample nucleic acids and the presence of sample nucleic acid may be determined by the presence or absence of label. This method provides reliable sample detection with about a 1000-fold reduction of molarity of the sample agent.

As a further aspect of the invention, the signal of the labelled probes may be amplified by means such as providing a common tail to the free probe which itself

comprises multiple chromogenic, enzymatic or radioactive labels or which is itself susceptible to specific binding by a further probe agent which is multiply labelled. In this way, a second round of signal amplification may be carried out. Labeled or unlabeled probes may be used in a second round of amplification. In this second round of amplification, a lengthy DNA sample with multiple labels may result in an increased amplification intensity signal between 10 to 100 fold which may result in a total signal amplification of 100,000 fold. Through the use of both aspects of this example, an intensity signal approximately 100,000 fold may give a positive result of probe-DNA ligation without having to employ PCR or other amplification procedures.

According to a further aspect of the invention an array or super array may be prepared which consists of a complete set of probes, for example 4096 6-mer probes. Arrays of this type are universal in a sense that they can be used for detection or partial to complete sequencing of any nucleic acid species. Individual spots in an array may contain single probe species or mixtures of probes, for example N(1-3) B(4-6) N(1-3) type of mixtures that are synthesized in the single reaction (N represents all four nucleotides, B one specific nucleotide and where the associated numbers are a range of numbers of bases *i.e.,* 1-3 means "from one to three bases"). These mixtures provide stronger signal for a nucleic acid species present at low concentration by collecting signal from different parts of the same long nucleic acid species molecule. The universal set of probes may be subdivided in many subsets which are spotted as unit arrays separated by barriers that prevent spreading of hybridization buffer with sample and labeled probe(s).

For detection of a nucleic acid species with a known sequence one of more oligonucleotide sequences comprising both unlabelled fixed and labeled probes in solution may be selected. Labeled probes are synthesized or selected from the presynthesized complete sets of, for example, 7-mers. The labeled probes are added to corresponding unit arrays of fixed probes such that a pair of fixed and labeled probes will adjacently hybridize to the target sequence such that upon administration of ligase the probes will be covalently bound.

If a unit array contains more than one fixed probe (as separated spots or within the same spot) that are positive in a given nucleic acid species all corresponding labeled probes may be mixed and added to the same unit array. The mixtures of labeled probes

are even more important when mixtures of nucleic acid species are tested. One example of a complex mixture of nucleic acid species are mRNAs in one cell or tissue.

According to one embodiment of the invention unit arrays of fixed probes allow use of every possible immobilized probe with cocktails of a relatively small number of labeled probes. More complex cocktails of labeled probes may be used if a multiplex labeling scheme is implemented. Preferred multiplexing methods may use different fluorescent dyes or molecular tags that may be separated by mass spectroscopy.

Alternatively, according to a preferred embodiment of the invention, relatively short fixed probes may be selected which frequently hybridize to many nucleic acid sequences. Such short probes are used in combination with a cocktail of labeled probes which may be prepared such that at least one labeled probe corresponds to each of the fixed proves. Preferred cocktails are those in which none of the labeled probes corresponds to more than one fixed probe.

5.45 INTERROGATION OF SEGMENTS OF THE HIV VIRUS WITH ALL POSSIBLE 5-MERS

In this example of Format III SBH, an array was generated on nylon membranes (*e.g.*, Gene Screen) of all possible bound 5-mers (1024 possible pentamers). The bound 5-mer oligonucleotides were synthesized with 5' tails of 5'-TTTTTT-NNN-3'(N = all four bases A, C, G, T, at this step in the synthesis equal molar amounts of all four bases are added). These oligonucleotides were precisely spotted onto the nylon membrane, the spots were allowed to dry, and the oligonucleotides were immobilized by treating the dried spots with UV light. Oligonucleotide densities of up to 18 oligonucleotides per square nanometer were obtained using this method. After the UV treatment, the nylon membranes were treated with a detergent containing buffer at 60-80°C. The spots of oligonucleotides were gridded in subarrays of 10 by 10 spots, and each subarray has 64 5-mer spots and 36 control spots. 16 subarrays give 1024 5-mers which encompasses all possible 5-mers.

The subarrays in the array were partitioned from each other by physical barriers, *e.g.*, a hydrophobic strip, that allowed each subarray to be hybridized to a sample without cross-contamination from adjacent subarrays. In a preferred embodiment, the

hydrophobic strip is made from a solution of silicone (*e.g.*, household silicone glue and seal paste) in an appropriate solvent (such solvents are well known in the art). This solution of silicone grease is applied between the subarrays to form lines which after the solvent evaporates act as hydrophobic strips separating the cells.

In this Format III example, the free or solution (nonbound) 5-mers were synthesized with 3' tails of 5'-NN-3' (N = all four bases A, C, G, T). In this embodiment, the free 5-mers and the bound 5-mers are combined to produce all possible I 0-mers for sequencing a known DNA sequence of less than 20kb. 20kb of double stranded DNA is denatured into 40 kb of single-stranded DNA. This 40 kb of ss DNA hybridizes to about 4% of all possible 10-mers. This low frequency of 10-mer binding and the known target sequence allow the pooling of free or solution (nonbound) 5-mers for treatment of each subarray, without a loss of sequence information. In a preferred embodiment, 16 probes are pooled for each subarray, and all possible 5-mers are represented in 64 total pools of free 5-mers. Thus, all possible I 0-mers may be probed against a DNA sample using 1024 subarrays (16 subarrays for each pool of free 5-mers).

The target DNA in this embodiment represents two-600 bp segments of the HIV virus. These 600 bp segments are represented by pools of 60 overlapping 30-mers (the 30-mers overlap each adjacent 30 mer by 20 nucleotides). The pools of 30-mers mimic a target DNA that has been treated using techniques well known in the art to shear, digest, and/or random PCR the target DNA to produce a random pool of very small fragments.

As described above in the previous Format III examples, the free 5-mers are labeled with radioactive isotopes, biotin, fluorescent dyes, etc. The labeled free 5-mers are then hybridized along with the bound 5-mers to the target DNA, and ligated. In a preferred embodiment, 300-1000 units of ligase are added to the reaction. The hybridization conditions were worked out following the teachings of the previous examples. Following ligation and removal of the target DNA and excess free probe, the array is assayed to determine the location of labeled probes (using the techniques described in the examples above).

The known DNA sequence of the target, and the known free and bound 5-mers in each subarray, predict which bound 5-mers will be ligated to a labeled free 5-mer in each subarray. The signal from 20 of these predicted dots were lost and 20 new signals were

gained for each change in the target DNA from the predicted sequence. The overlapping sequence of the bound 5-mers in these ten new dots identifies which free, labeled 5-mer is bound in each new dot.

Using the described methods, arrays and pools of free, labeled 5-mers, the test HIV DNA sequence was probed with all possible 10mers. Using this Format III approach, we properly identified the "wild-type" sequence of the segments tested, as well as several sequence "mutants" that were introduced into these segments.

5.46 SEQUENCING OF REPETITIVE DNA SEQUENCES

Repetitive DNA sequences in the target DNA are sequenced with "spacer oligonucleotides" in a modified Format III approach. Spacer oligonucleotides of varying lengths of the repetitive DNA sequence (the repeating sequence is identified on a first SBH run) are hybridized to the target DNA along with a first known adjoining oligonucleotide and a second known, or group of possible oligonucleotides adjoining the other side of the spacer (known from the first SBH run). When a spacer matching the length of the repetitive DNA segment is hybridized to the target, the two adjacent oligonucleotides can be ligated to the spacer. If the first known oligonucleotide is fixed to a substrate, and the second known or possible oligonucleotide(s)is labeled, a bound ligation product including the labeled second known or possible oligonucleotide(s)is formed when a spacer of the proper length is hybridized to the target DNA.

5.47 SEQUENCING THROUGH BRANCH POINTS WITH FORMAT III SBH

Branch points in the target DNA are sequenced using a third set of oligonucleotides and a modified Format III approach. After a first SBH run, several branch points may be identified when the sequence is compiled. These can be solved by hybridizing oligonucleotide(s)that overlap partially with one of the known sequences leading into the branch point and then hybridizing to the target an additional oligonucleotide that is labeled and corresponds to one of the sequences that comes out of the branch point. When the proper oligonucleotides are hybridized to the target DNA, the labeled oligonucleotide can be ligated to the other(s). In a preferred embodiment, a first

oligonucleotide that is offset by one to several nucleotides from the branch point is selected (so that it reads into one of the branch sequences), a second oligonucleotide reading from the first and into the branch point sequence is also selected, and a set of third oligonucleotides that correspond to all the possible branch sequences with an overlap of the branch point sequence by one or a few nucleotides (corresponding to the first oligonucleotide) is selected. These oligonucleotides are hybridized to the target DNA, and only the third oligonucleotide with the proper branch sequence (that matches the branch sequence of the first oligonucleotide) will produce a ligation product with the first and second oligonucleotides.

5.48 MULTIPLEXING PROBES FOR ANALYZING A TARGET NUCLEIC ACID

Sets of probes are labeled with different labels so that each probe of a set can be differentiated from the other probes in the set. Thus, the set of probes may be contacted with target nucleic acid in a single hybridization reaction without the loss of any probe information. In preferred embodiments, the different labels are different radioisotopes, or different fluorescent labels, or different EMLs. These sets of probes may be used in either Format I, Format II or Format III SBH.

In Format 1 SBH, the set of differently labeled probes are hybridized to target nucleic acid which is fixed to a substrate under conditions that allow differentiation between perfect matches one base-pair mismatches. Specific probes which bind to the target nucleic acid are identified by their different labels and perfect matches are determined, at least in part, from this binding information.

In Format II SBH, the target nucleic acids are labeled with different probes and hybridized to arrays of probes. Specific target nucleic acids which bind to the probes are identified by their different labels and perfect matches are determined, at least in part, form this binding information.

In Format III SBH, the set of differently labeled probes and fixed probes are hybridized to a target nucleic acid under conditions that allow perfect matches to be differentiated from one base-pair mismatches. Labeled probes that are adjacent on the

target to a fixed probe are bound to the fixed probe, and these products are detected and differentiated by their different labels.

In a preferred embodiment, the different labels are EMLs, which can be detected by electron capture mass spectrometry (EC-MS). EMLs may be prepared from a variety of backbone molecules, with certain aromatic backbones being particularly preferred, *e.g.,* see Xu *et al..*, J. Chromatog. 764:95-102 (1997). The EML is attached to a probe in a reversible and stable manner, and after the rpobe is hybridized to target nucleic acid, the EML is removed from the probe and identified by standard EC-MS (*e.g.,* the EC-MS may be done by a gas chromotograph-massspectrometer).

5.49 DETECTION OF LOW FREQUENCY TARGET NUCLEIC ACIDS

Format III SBH has sufficient discrimination power to identify a sequence that is present in a sample at 1 part to 99 parts of a similar sequence that differs by a single nucleotide. Thus, Format III can be used to identify a nucleic acid present at a very low concentration in a sample of nucleic acids, *e.g.,* a sample derived from blood.

In one embodiment, the two sequences are for cystic fibrosis and the sequences differ from each other by a deletion of three nucleotides. Probes for the two sequences were as follows, probes distinguishing the deletion from wild type were fixed to a substrate, and a labeled contiguous probe was common to both. Using these targets and probes, the deletion mutant could be detected with Format III SBH when it was present at one part to ninety nine pans of the wild-type.

5.50 POLAROID APPARATUS AND METHOD FOR ANALYZING A TARGET NUCLEIC ACID

An apparatus for analyzing a nucleic acid can be constructed with two arrays of nucleic acids, and an optional material that prevents the nucleic acids of the two arrays from mixing until such mixing is desired. The arrays of the apparatus may be supported by a variety of substrates, including but not limited to, nylon membranes, nitrocellulose membranes, or other materials disclosed above. In preferred embodiments, one of the substrate is a membrane separated into sectors by hydrophobic strips, or a suitable support material with wells which may contain a gel or sponge. In this embodiment,

probes are placed on a sector of the membrane, or in the well, the gel, or sponge, and a solution (with or without target nucleic acids) is added to the membrane or well so that the probes are solubilized. The solution with the solubilized probes is then allowed to contact the second array of nucleic acids. The nucleic acids may be, but are not limited to, oligonucleotide probes, or target nucleic acids, and the probes or target nucleic acids may be labeled. The nucleic acids may be labeled with any labels conventionally used in the art, including but not limited to radioisotopes, fluorescent labels or electrophore mass labels.

The material which prevents mixing of the nucleic acids may be disposed between the two arrays in such a way that when the material is removed the nucleic acids of the two arrays mix together. This material may be in the form of a sheet, membrane, or other barrier, and this material may be comprised of any material that prevents mixing of the nucleic acids.

This apparatus may be used in Format I SBH as follows: a first array of the apparattus has target nucleic acids that are fixed to the substrate, and a second array of the apparatus has nucleic acid probes that are labeled and can be removed to interrogate the target nucleic acid of the first array. The two arrays are optionally separated by a sheet of material that prevents the probes from contacting the target nucleic acid, and when this sheet is removed, the probes can interrogate the target. After appropriate incubation, and (optionally) washing steps the array of targets may be read to determine which probes formed perfect matches with the target. This reading may be automated or can be done manually (e.g. by eye with an autoradiogram). In format II SBH, the procedure followed would be similar to that described above except the target is labeled and the probes are fixed.

Alternatively the apparatus may be used in Format III SBH as follows: to arrays of nucleic acid probes are formed, the nucleic acid probes of either or both arrays may be labeled and one of the arrays may be fixed to its substarate. The two arrays are separated by a sheet of material that prevents the probes from mixing. A Format II reaction is initiated by adding target nucleic acid and removing the sheet allowing the probes to mix with each other and the target. Probes which bind to adjacent sites on the target are bound together (*e.g.*, by base stacking interactions or by covalently joining the

backbones), and the results are read to determine which probes bound to the target at adjacent sites. When one set of probes is fixed to the substrate, the fixed array can be read to determine which probes from the other array are bound together with the fixed probes. As with the method above, this reading may be automated (*e.g.*, with an ELISA reader) or can be done manually (*e.g.*, by eye with an autoradiogram).

6.0 EXAMPLES

6.1 The 748 Gene Family

Novel Contigs

The novel contigs of the invention, were assembled from novel expressed sequence tags (EST's) isolated by methods described herein *(e.g.*, SBH), and in some cases sequences obtained from one or more public databases. The inserts for the cDNA libraries from which the novel ESTs were obtained were amplified with PCR using primers specific for the vector sequences which flank the inserts. These samples were spotted onto nylon membranes and interrogated with oligonucleotide probes to give sequence signatures. The clones were clustered into groups of similar or identical sequences, and single representative clones were selected from each group for gel sequencing. The 5□sequence of the amplified inserts was then deduced using the reverse M13 sequencing primer in a typical Sanger sequencing protocol. PCR products were purified and subjected to fluorescent dye terminator cycle sequencing. Single pass gel sequencing was done using a 377 Applied Biosystems (ABI) sequencer to obtain the novel ESTs. The novel contigs of the invention, were assembled from the novel ESTs and, in some cases, sequences obtained from one or more public databases. The sequences for the resulting contigs from the 748 gene family are designated as 748 SEQ ID NO: 1-45,207 and are provided in the Sequence Listing with corresponding SEQ ID NOs 1-45,207.

6.2 The 752 Gene Family

Novel Contigs

The novel contigs of the invention, were assembled from novel expressed sequence tags (EST's) isolated by methods described herein *(e.g.*, SBH), and in some

cases sequences obtained from one or more public databases. The inserts for the cDNA libraries from which the novel ESTs were obtained were amplified with PCR using primers specific for the vector sequences which flank the inserts. These samples were spotted onto nylon membranes and interrogated with oligonucleotide probes to give sequence signatures. The clones were clustered into groups of similar or identical sequences, and single representative clones were selected from each group for gel sequencing. The 5□sequence of the amplified inserts was then deduced using the reverse M13 sequencing primer in a typical Sanger sequencing protocol. PCR products were purified and subjected to fluorescent dye terminator cycle sequencing. Single pass gel sequencing was done using a 377 Applied Biosystems (ABI) sequencer to obtain the novel ESTs. The novel contigs of the invention, were assembled from the novel ESTs and, in some cases, sequences obtained from one or more public databases. The sequences for the resulting contigs from the 752 gene family are designated as 752 SEQ ID NO: 1-13,203 and are provided in the Sequence Listing with the corresponding SEQ ID NOs 45,208-58,410.

6.3 The 778 Gene Family

Novel Contigs

The novel contigs of the invention, were assembled from novel expressed sequence tags (ESTs) isolated by methods described herein *(e.g.,* SBH), and in some cases sequences obtained from one or more public databases. The inserts for the cDNA libraries from which the novel ESTs were obtained were amplified with PCR using primers specific for the vector sequences which flank the inserts. These samples were spotted onto nylon membranes and interrogated with oligonucleotide probes to give sequence signatures. The clones were clustered into groups of similar or identical sequences, and single representative clones were selected from each group for gel sequencing. The 5□sequence of the amplified inserts was then deduced using the reverse Ml3 sequencing primer in a typical Sanger sequencing protocol. PCR products were purified and subjected to fluorescent dye terminator cycle sequencing. Single pass gel sequencing was done using a 377 Applied Biosystems (ABI) sequencer to obtain the novel ESTs. The novel contigs of the invention, were assembled from the novel ESTs

and, in some cases, sequences obtained from one or more public databases. The sequences for the resulting contigs from the 778 gene family are designated as 778 SEQ ID NO: 1-105 and are provided in the Sequence Listing with the corresponding SEQ ID NOs 58,411-58,515.

5

6.4 The 779 Gene Family

Novel Contigs

The novel contigs of the invention, were assembled from sequences that were obtained from cDNA libraries by methods described herein *(e.g.,* SBH). Briefly, clones
10 from cDNA libraries were spotted on nylon membrane filters and screened with oligonucleotide probes (*e.g.,* 7-mers) to obtain signature sequences. The clones were clustered into groups of similar or identical sequences, and single representative clones were selected from each group for gel sequencing. The inserts for the cDNA libraries from which the sequences were obtained were amplified with PCR using primers specific
15 for the vector sequences which flank the inserts, or isolated from plasmid preparations. The 5' sequence of the amplified inserts was then deduced using the reverse M13 sequencing primer in a typical Sanger sequencing protocol, as well as internal primers in both the forward and reverse direction. In some cases RACE (Random Amplification of cDNA Ends) was performed to further extend the sequence in the 5□direction. In some
20 cases all of a signature cluster was sequenced to generate overlapping clones to assemble the contigs.

Chromatograms were base called and assembled using a software suite from Washington University, MO containing three applications designated PHRED, PEIRAP, and CONSED. The sequences for the resulting contigs for the 779 gene family are
25 designated as 779 SEQ ID NO: 1-128 and are provided in the Sequence Listing with the corresponding SEQ ID NOs 58,516-58,643.

Additionally, contig sequences were BLASTed against Hyseq's database to to determine adequate sequence homology for addition to the contigs. Full length clones of the entire message were obtained either by identifying clones which contained the
30 beginning of the open reading frame and fully sequenced and verified against the contig. If no clones were available containing the full length sequence, PCR was used to generate

an amplicon from tissue libraries and the entire length sequenced and verified against the contig.

6.5 **The 782 Gene Family**

Novel Contigs

A novel contig of the invention was assembled from sequences that were obtained from a cDNA library by methods described herein (*e.g.,* SBH). Briefly, clones from cDNA libraries were spotted on nylon membrane filters and screened with oligonucleotide probes (*e.g.,* 7-mers) to obtain signature sequences. The inserts for the cDNA libraries from which the sequences were obtained were amplified with PCR using primers specific for the vector sequences which flank the inserts, or isolated from plasmid preparations. The 5' sequence of the amplified inserts was then deduced using the reverse M13 sequencing primer in a typical Sanger sequencing protocol, as well as internal primers in both the forward and reverse direction. In some cases RACE (Random Amplification of cDNA Ends) was performed to further extend the sequence in the 5' direction. In all cases all of a signature cluster was sequenced to generate overlapping clones to assemble the contigs. Chromatograms were base called and assembled using a software suite from University of Washington, Seattle containing three applications designated PHRED, PHRAP, and CONSED. The sequences for the resulting contigs for the 782 gene family are designated as 782 SEQ ID NO: 1-10,451 and are provided in the attached Sequence Listing with the corresponding SEQ ID NOs 58,644-69,094. inserts was then deduced in a typical Sanger sequencing protocol. The inserts of the library were, amplified with PCR using 5 primers specific for vector sequences which flank the inserts.

The contigs were assembled using an EST sequence as a seed. Then a recursive algorithm was used to extend the seed EST into an extended assemblage, by pulling additional sequences from different databases (*i.e.,* Hyseq's database containing EST sequences, dbEST version 114, gb pri 114, and UniGene version 101) that belong to this assemblage. The algorithm terminated when there was no additional sequences from the above databases that would extend the assemblage. Inclusion of component sequences into the assemblage was based on a BLASTN hit to the extending assemblage with BLAST score greater than 300 and percent identity greater than 95%.

The nearest neighbor result for the assembled contig was obtained by a FASTA version 3 search against Genpept release 114, using FASTXY algorithm. FASTXY is an improved version of FASTA alignment which allows in-codon frame shifts. The nearest neighbor result showed the closest homologue for each assemblage from Genpept (and contains the translated amino acid sequences for which the assemblage encodes).

6.6 The 784 Gene Family

Novel Contigs

Novel predicted polypeptides (including proteins) encoded by the novel polynucleotides (784 SEQ ID NO: 1-10,289) of the present invention, and their corresponding nucleotide locations to each of 748 SEQ ID NO: 1-10,289 were determined. Methods A, B, and C were used to predict the polypeptides. Method A refers to a polypeptide obtained by using a software program called FASTY (available from http://fasta.bioch.virginia.edu) which selects a polypeptide based on a comparison of translated novel polynucleotide to known polypeptides (W.R. Pearson, Methods in Enzymology, 183: 63-98 (1990), incorporated herein by reference). Method B refers to a polypeptide obtained by using a software program called GenScan for human/vertebrate sequences (available from Stanford University, Office of Technology Licensing) that predicts the polypeptide based on a probabilistic model of gene structure/compositional properties (C. Burge and S. Karlin, J. Mol. Biol., 268: 78-94 (1997), incorporated herein by reference). Method C refers to a polypeptide obtained by using a Hyseq proprietary software program that translates the novel polynucleotide and its complementary strand into six possible amino acid sequences (forward and reverse frames) and chooses the polypeptide with the longest open reading frame. When the predicted beginning nucleotide is a higher number than the predicted end nucleotide, then the amino acid sequence is derived from the complementary strand of the indicated SEQ ID NO. The locations of the predicted beginning and end nucleotides correlate to the nucleotide sequence of the indicated SEQ ID NO., not its complementary strand.

The isolated polypeptides of the invention include, but are not limited to, a polypeptide comprising any of the amino acid sequences predicted by the methods described above or from six frame translations of 784 SEQ ID NO: 1-10,289; or the

corresponding full length or mature protein. One of skill in the art could determine the corresponding amino acid sequence using techniques well known in the art to translate and analyze all possible six frames. Polypeptides of the invention also include polypeptides with biological activity that are encoded by (a) any of the polynucleotides having a nucleotide sequence set forth in the 784 SEQ ID NO: 1-10,289; or (b) polynucleotides that hybridize to the complement of the polynucleotids of (a) under stringent hybridization conditions. Biologically or immunologically active variants of any of the polypeptide sequences or from six frame translations of 784 SEQ ID NO: 1-10,289, and "substantial equivalents" thereof (*e.g.*, with at least about 65%, 70%, 75%, 80%, 85%, 90%, 95%, 98% or 99% amino acid sequence identity) that preferably retain biological activity are also contemplated. The polypeptides of the invention may be wholly or partially chemically synthesized but are preferably produced by recombinant means using the genetically engineered cells (e.g. host cells) of the invention.

A plurality of novel nucleic acids were obtained from cDNA libraries prepared from various human tissues and in some cases isolated from a genomic library derived from human chromosome using standard PCR, SBH sequence signature analysis and Sanger sequencing techniques. The inserts of the library were amplified with PCR using primers specific for the vector sequences which flank the inserts. Clones from cDNA libraries were spotted on nylon membrane filters and screened with oligonucleotide probes (*e.g.*, 7-mers) to obtain signature sequences. The clones were clustered into groups of similar or identical sequences. Representative clones were selected for sequencing.

In some cases, the 5' sequence of the amplified inserts was then deduced using a typical Sanger sequencing protocol. PCR products were purified and subjected to fluorescent dye terminator cycle sequencing. Single pass gel sequencing was done using a 377 Applied Biosystems (ABI) sequencer to obtain the novel nucleic acid sequences. In some cases RACE (Random Amplification of cDNA Ends) was performed to further extend the sequence in the 5' direction.

The contigs or the nucleic acids of the present invention, designated as 784 SEQ ID NO: 1-10,289 were assembled using an EST sequence as a seed. Then a recursive algorithm was used to extend the seed EST into an extended assemblage, by pulling additional sequences from different databases (*i.e.*, Hyseq's database containing EST

sequences, dbEST version 114, gb pri 114, and UniGene version 101) that belong to this assemblage. The algorithm terminated when there was no additional sequences from the above databases that would extend the assemblage. Inclusion of component sequences into the assemblage was based on a BLASTN hit to the extending assemblage with BLAST score greater than 300 and percent identity greater than 95%.

The nearest neighbor result for the assembled contig was obtained by a FASTA version 3 search against Genpept release 114, using FASTXY algorithm. FASTXY is an improved version of FASTA alignment which allows in-codon frame shifts. The nearest neighbor result showed the closest homologue for each assemblage from Genpept (and contains the translated amino acid sequences for which the assemblage encodes).

6.7 **The 785 Gene Family**

Novel Nucleic Acid Sequences Obtained From Various Libraries

A plurality of novel nucleic acids were obtained from cDNA libraries prepared from various human tissues and in some cases isolated from a genomic library derived from human chromosome using standard PCR, SBH sequence signature analysis and Sanger sequencing techniques. The inserts of the library were amplified with PCR using primers specific for the vector sequences which flank the inserts. Clones from cDNA libraries were spotted on nylon membrane filters and screened with oligonucleotide probes (*e.g.*, 7-mers) to obtain signature sequences. The clones were clustered into groups of similar or identical sequences. Representative clones were selected for sequencing.

In some cases, the 5' sequence of the amplified inserts was then deduced using a typical Sanger sequencing protocol. PCR products were purified and subjected to fluorescent dye terminator cycle sequencing. Single pass gel sequencing was done using a 377 Applied Biosystems (ABI) sequencer to obtain the novel nucleic acid sequences. In some cases RACE (Random Amplification of cDNA Ends) was performed to further extend the sequence in the 5' direction.

The novel contigs of the invention were assembled from sequences that were obtained from a cDNA library by methods described above, and in some cases sequences obtained from one or more public databases. Chromatograms were base called and assembled using a software suite from University of Washington, Seattle containing three

applications designated PJRED, PHRAP, and CONSED. The sequences for the resulting contigs are designated as 785 SEQ ID NO: 1-3,796 and are provided in the Sequence Listing with the corresponding SEQ ID NOs 79,384-83,179. The contigs were assembled using an EST sequence as a seed. Then a recursive algorithm was used to extend the seed EST into an extended assemblage, by pulling additional sequences from different databases (*i.e.,* Hyseq's database containing EST sequences, dbEST version 114, gb pri 114, and UniGene version 101) that belong to this assemblage. The algorithm terminated when there was no additional sequences from the above databases that would extend the assemblage. Inclusion of component sequences into the assemblage was based on a BLASTN hit to the extending assemblage with BLAST score greater than 300 and percent identity greater than 95%.

The nearest neighbor result for the assembled contig was obtained by a FASTA version 3 search against Genpept release 114, using Fastxy algorithm. Fastxy is an improved version of FASTA alignment which allows in-codon frame shifts. The nearest neighbor result showed the closest homologue for each assemblage from Genpept (and contains the translated amino acid sequences for which the assemblage encodes). The nucleotide sequence within the assembled contigs that codes for signal peptide sequences and their cleavage sites can be determined from using Neural network SignalP V1.1 program (from Center for Biological Sequence Analysis, The Technical University of Denmark). The process for identifying prokaryotic and eukaryotic signal peptidesand their cleavage sites are also disclosed iby Henrick Nielson, Jacob Englebrecht, Soren Brunak, and Gunnar von Heijne in the publication "Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites" Protein Engineering, vol. 10, no. 1, pp. 1-6 (1997) incorporated herein by reference. A maximum S score and a mean S score, as described in the Nielson et. al., reference, are obtained from each assembled contig. The nucleotide sequence range for each sequence of 785 SEQ ID NO: 1-3,796 that encodes a corresponding forty-five amino acid sequence containing the signal peptide sequence and its cleavage site, the maximum S score and the mean S score obtained for each sequence were determined. Not all forty-five amino acids in the sequence may comprise the signal peptide.

6.8 **The 787 Gene Family**

The novel predicted polypeptides (including proteins) encoded by the novel polynucleotides (787 SEQ ID NO: 1-10,410) of the present invention, and their corresponding nucleotide locations to each of 787 SEQ ID NO: 1-10,410 were determined. Methods A, B, and C were used to predict the polypeptides. Method A refers to a polypeptide obtained by using a software program called FASTY (available from http://fasta.bioch.virginia.edu) which selects a polypeptide based on a comparison of translated novel polynucleotide to known polypeptides (W.R. Pearson, Methods in Enzymology, 183: 63-98 (1990), incorporated herein by reference). Method B refers to a polypeptide obtained by using a software program called GenScan for human/vertebrate sequences (available from Stanford University, Office of Technology Licensing) that predicts the polypeptide based on a probabilistic model of gene structure/compositional properties (C. Burge and S. Karlin, J. Mol. Biol., 268: 78-94 (1997), incorporated herein by reference). Method C refers to a polypeptide obtained by using a Hyseq proprietary software program that translates the novel polynucleotide and its complementary strand into six possible amino acid sequences (forward and reverse frames) and chooses the polypeptide with the longest open reading frame. When the predicted beginning nucleotide of is a higher number than the predicted end nucleotide, then the amino acid sequence is derived from the complementary strand of the indicated SEQ ID NO. The locations of the predicted beginning and end nucleotides correlate to the nucleotide sequence of the indicated SEQ ID NO., not its complementary strand.

The isolated polypeptides of the invention include, but are not limited to, a polypeptide comprising any of the predicted amino acid sequences or from six frame translations of 787 SEQ ID NO: 1-10,410; or the corresponding full length or mature protein. One of skill in the art could determine the corresponding amino acid sequence using techniques well known in the art to translate and analyze all possible six frames. Polypeptides of the invention also include polypeptides with biological activity that are encoded by (a) any of the polynucleotides having a nucleotide sequence set forth in the 787 SEQ ID NO: 1-10,410; or (b) polynucleotides that hybridize to the complement of the polynucleotids of (a) under stringent hybridization conditions. Biologically or immunologically active variants of any of the polypeptide sequences or from six frame

translations of 787 SEQ ID NO: 1-10,410, and "substantial equivalents" thereof (*e.g.*, with at least about 65%, 70%, 75%, 80%, 85%, 90%, 95%, 98% or 99% amino acid sequence identity) that preferably retain biological activity are also contemplated. The polypeptides of the invention may be wholly or partially chemically synthesized but are preferably produced by recombinant means using the genetically engineered cells (e.g. host cells) of the invention.

A plurality of novel nucleic acids were obtained from cDNA libraries prepared from various human tissues and in some cases isolated from a genomic library derived from human chromosome using standard PCR, SBH sequence signature analysis and Sanger sequencing techniques. The inserts of the library were amplified with PCR using primers specific for the vector sequences which flank the inserts. Clones from cDNA libraries were spotted on nylon membrane filters and screened with oligonucleotide probes (*e.g.*, 7-mers) to obtain signature sequences. The clones were clustered into groups of similar or identical sequences. Representative clones were selected for sequencing.

In some cases, the 5' sequence of the amplified inserts was then deduced using a typical Sanger sequencing protocol. PCR products were purified and subjected to fluorescent dye terminator cycle sequencing. Single pass gel sequencing was done using a 377 Applied Biosystems (ABI) sequencer to obtain the novel nucleic acid sequences. In some cases RACE (Random Amplification of cDNA Ends) was performed to further extend the sequence in the 5' direction.

Novel Contigs

The contigs or the nucleic acids of the present invention, designated as 787 SEQ ID NO: 1-10,410 were assembled using an EST sequence as a seed. Then a recursive algorithm was used to extend the seed EST into an extended assemblage, by pulling additional sequences from different databases (*i.e.*, Hyseq's database containing EST sequences, dbEST version 114, gb pri 114, and UniGene version 101) that belong to this assemblage. The algorithm terminated when there was no additional sequences from the above databases that would extend the assemblage. Inclusion of component sequences into the assemblage was based on a BLASTN hit to the extending assemblage with BLAST score greater than 300 and percent identity greater than 95%.

The nearest neighbor result for the assembled contig was obtained by a FASTA version 3 search against Genpept release 114, using FASTXY algorithm. FASTXY is an improved version of FASTA alignment which allows in-codon frame shifts. The nearest neighbor result showed the closest homologue for each assemblage from Genpept (and contains the translated amino acid sequences for which the assemblage encodes).

6.9 The 788 Gene Family

The novel predicted polypeptides (including proteins) encoded by the novel polynucleotides (788 SEQ ID NO: 1-14,074) of the present invention, and their corresponding nucleotide locations to each of 788 SEQ ID NO: 1-14,074 were determined. Methods A, B, and C were used to predict the polypeptides. Method A refers to a polypeptide obtained by using a software program called FASTY (available from http://fasta.bioch.virginia.edu) which selects a polypeptide based on a comparison of translated novel polynucleotide to known polypeptides (W.R. Pearson, Methods in Enzymology, 183: 63-98 (1990), incorporated herein by reference). Method B refers to a polypeptide obtained by using a software program called GenScan for human/vertebrate sequences (available from Stanford University, Office of Technology Licensing) that predicts the polypeptide based on a probabilistic model of gene structure/compositional properties (C. Burge and S. Karlin, J. Mol. Biol., 268: 78-94 (1997), incorporated herein by reference). Method C refers to a polypeptide obtained by using a Hyseq proprietary software program that translates the novel polynucleotide and its complementary strand into six possible amino acid sequences (forward and reverse frames) and chooses the polypeptide with the longest open reading frame. When the predicted beginning nucleotide is a higher number than the predicted end nucleotide, then the amino acid sequence is derived from the complementary strand of the indicated SEQ ID NO. The locations of the predicted beginning and end nucleotides correlate to the nucleotide sequence of the indicated SEQ ID NO., not its complementary strand.

The isolated polypeptides of the invention include, but are not limited to, a polypeptide comprising any of the amino acid sequences or from six frame translations of 788 SEQ ID NO: 1-14,074; or the corresponding full length or mature protein. One of skill in the art could determine the corresponding amino acid sequence using techniques

well known in the art to translate and analyze all possible six frames. Polypeptides of the invention also include polypeptides with biological activity that are encoded by (a) any of the polynucleotides having a nucleotide sequence set forth in the 788 SEQ ID NO: 1-14,074; or (b) polynucleotides that hybridize to the complement of the polynucleotides of (a) under stringent hybridization conditions. Biologically or immunologically active variants of any of the polypeptide sequences or from six frame translations of 788 SEQ ID NO: 1-14,074, and "substantial equivalents" thereof (*e.g.*, with at least about 65%, 70%, 75%, 80%, 85%, 90%, 95%, 98% or 99% amino acid sequence identity) that preferably retain biological activity are also contemplated. The polypeptides of the invention may be wholly or partially chemically synthesized but are preferably produced by recombinant means using the genetically engineered cells (e.g. host cells) of the invention.

A plurality of novel nucleic acids were obtained from cDNA libraries prepared from various human tissues and in some cases isolated from a genomic library derived from human chromosome using standard PCR, SBH sequence signature analysis and Sanger sequencing techniques. The inserts of the library were amplified with PCR using primers specific for the vector sequences which flank the inserts. Clones from cDNA libraries were spotted on nylon membrane filters and screened with oligonucleotide probes (*e.g.*, 7-mers) to obtain signature sequences. The clones were clustered into groups of similar or identical sequences. Representative clones were selected for sequencing. In some cases the 5' sequence of the amplified inserts was then deduced using a typical Sanger sequencing protocol. PCR products were purified and subjected to fluorescent dye terminator cycle sequencing. Single pass gel sequencing was done using a 377 Applied Biosystems (ABI) sequencer to obtain the novel nucleic acid sequences. In some cases RACE (Random Amplification of cDNA Ends) was performed to further extend the sequence in the 5' direction.

Novel Contigs

The novel contigs of the invention were assembled from sequences that were obtained from a cDNA library by methods described above, and in some cases sequences obtained from one or more public databases. The sequences for the resulting contigs are designated as 788 SEQ ID NO: 1-14,074 and are provided in the attached Sequence

Listing with the corresponding SEQ ID NOs 93,590-107,663; The contigs were assembled using an EST sequence as a seed. Then a recursive algorithm was used to extend the seed EST into an extended assemblage, by pulling additional sequences from different databases (*i.e.,* Hyseq's database containing EST sequences, dbEST version 114, gb pri 114, and UniGene version 101) that belong to this assemblage. The algorithm terminated when there was no additional sequences from the above databases that would extend the assemblage. Inclusion of component sequences into the assemblage was based on a BLASTN hit to the extending assemblage with BLAST score greater than 300 and percent identity greater than 95%.

The nearest neighbor result for the assembled contig was obtained by a FASTA version 3 search against Genpept release 115, using FASTXY algorithm. FASTXY is an improved version of FASTA alignment which allows in-codon frame shifts. The nearest neighbor result showed the closest homologue for each assemblage from Genpept (and contains the translated amino acid sequences for which the assemblage encodes).

6.10 The 789 Gene Family

The novel predicted polypeptides (including proteins) encoded by the novel polynucleotides (789 SEQ ID NO: 1-6,391) of the present invention, and their corresponding nucleotide locations to each of SEQ ID NO: 1-6,391 were determined. Methods A, B, and C were used to predict the polypeptides. Method A refers to a polypeptide obtained by using a software program called FASTY (available from http://fasta.bioch.virginia.edu) which selects a polypeptide based on a comparison of translated novel polynucleotide to known polypeptides (W.R. Pearson, Methods in Enzymology, 183: 63-98 (1990), incorporated herein by reference). Method B refers to a polypeptide obtained by using a software program called GenScan for human/vertebrate sequences (available from Stanford University, Office of Technology Licensing) that predicts the polypeptide based on a probabilistic model of gene structure/compositional properties (C. Burge and S. Karlin, J. Mol. Biol., 268: 78-94 (1997), incorporated herein by reference). Method C refers to a polypeptide obtained by using a Hyseq proprietary software program that translates the novel polynucleotide and its complementary strand

into six possible amino acid sequences (forward and reverse frames) and chooses the polypeptide with the longest open reading frame.

A plurality of novel nucleic acids were obtained from cDNA libraries prepared from various human tissues and in some cases isolated from a genomic library derived from human chromosome using standard PCR, SBH sequence signature analysis and Sanger sequencing techniques. The inserts of the library were amplified with PCR using primers specific for the vector sequences which flank the inserts. Clones from cDNA libraries were spotted on nylon membrane filters and screened with oligonucleotide probes (*e.g.*, 7-mers) to obtain signature sequences. The clones were clustered into groups of similar or identical sequences. Representative clones were selected for sequencing.

In some cases, the 5' sequence of the amplified inserts was then deduced using a typical Sanger sequencing protocol. PCR products were purified and subjected to fluorescent dye terminator cycle sequencing. Single pass gel sequencing was done using a 377 Applied Biosystems (ABI) sequencer to obtain the novel nucleic acid sequences. In some cases RACE (Random Amplification of cDNA Ends) was performed to further extend the sequence in the 5' direction.

Novel Contigs

The novel contigs or the nucleic acids of the present invention of the invention were assembled from sequences that were obtained from a cDNA library by methods described above, and in some cases sequences obtained from one or more public databases. The sequences for the resulting contigs are designated as 789 SEQ ID NO: 1-6,391 and are provided in the attached Sequence Listing with the corresponding SEQ ID NOs 107,664-114,054. The contigs were assembled using an EST sequence as a seed. Then a recursive algorithm was used to extend the seed EST into an extended assemblage, by pulling additional sequences from different databases (*i.e.*, Hyseq's database containing EST sequences, dbEST version 114, gb pri 114, and UriGene version 101) that belong to this assemblage. The algorithm terminated when there was no additional sequences from the above databases that would extend the assemblage. Inclusion of component sequences into the assemblage was based on a BLASTN hit to the extending assemblage with BLAST score greater than 300 and percent identity greater than 95 %.

The nearest neighbor result far the assembled contig was obtained by a FASTA version 3 search against Genpept release 115, using FASTXY algorithm. FASTXY is an improved version of FASTA alignment which allows in-codon frame shifts. The nearest neighbor result showed the closest homologue for each assemblage from Genpept (and contains the translated amino acid sequences for which the assemblage encodes

6.11 The 790 Gene Family

The novel predicted polypeptides (including proteins) encoded by the novel polynucleotides (790 SEQ ID NO: 1-30,553) of the present invention, and their corresponding start and stop nucleotide location to each of 790 SEQ ID NO: 1-30,553. Methods A, B, and C were used to predict the polypeptides. Method A refers to a polypeptide obtained by using a software program called FASTY (available from http://fasta.bioch.virginia.edu) which selects a polypeptide based on a comparison of translated novel polynucleotide to known polypeptides (W.R. Pearson, Methods in Enzymology, 183: 63-98 (1990), incorporated herein by reference). Method B refers to a polypeptide obtained by using a software program called GenScan for human/vertebrate sequences (available from Stanford University, Office of Technology Licensing) that predicts the polypeptide based on a probabilistic model of gene structure/compositional properties (C. Burge and S. Karlin, J. Mol. Biol., 268: 78-94 (1997), incorporated herein by reference). Method C refers to a polypeptide obtained by using a Hyseq proprietary software program that translates the novel polynucleotide and its complementary strand into six possible amino acid sequences (forward and reverse frames) and chooses the polypeptide with the longest open reading frame.

A plurality of novel nucleic acids were obtained from cDNA libraries prepared from various human tissues and in some cases isolated from a genomic library derived from human chromosome using standard PCR, SBH sequence signature analysis and Sanger sequencing techniques. The inserts of the library were amplified with PCR using primers specific for the vector sequences which flank the inserts. Clones from cDNA libraries were spotted on nylon membrane filters and screened with oligonucleotide probes (*e.g.*, 7-mers) to obtain signature sequences. The clones were clustered into groups of similar or identical sequences. Representative clones were selected for sequencing.

In some cases, the 5' sequence of the amplified inserts was then deduced using a typical Sanger sequencing protocol. PCR products were purified and subjected to fluorescent dye terminator cycle sequencing. Single pass gel sequencing was done using a 377 Applied Biosystems (ABI) sequencer to obtain the novel nucleic acid sequences. In some cases RACE (Random Amplification of cDNA Ends) was performed to further extend the sequence in the 5' direction.

Novel Contigs

The contigs or the nucleic acids of the present invention, designated as 790 SEQ ID NO: 1-30,553 were assembled using an EST sequence from Hyseq's database as a seed. A recursive algorithm was used to extend the seed EST into an extended assemblage, by pulling additional sequences from different databases (*e.g.,* Hyseq's database containing EST sequences, dbEST version 115, gb pri 115, and UniGene version 10.3, and exons from public domain genomic sequences predicted by GenScan) that belong to this assemblage. The algorithm terminated when there were no additional sequences from the databases that will extend the assemblage. Further, the inclusion of component sequences into the assemblage was based on a BLASTN hit to the extending assemblage with BLAST score greater than 300 and percent identity greater than 95%.

The nearest neighbor result for the assembled contig was obtained by a FASTA version 3 search against Genpept release 1.15, using FASTXY algorithm. FASTXY is an improved version of FASTA alignment which allows in-codon frame shifts. The nearest neighbor result showed the closest homologue for each assemblage from Genpept (and contains the translated amino acid sequences for which the assemblage encodes).

6.12 The 791 Gene Family

The novel predicted polypeptides (including proteins) encoded by the novel polynucleotides (791 SEQ ID NO: 1-5,822) of the present invention, and their corresponding nucleotide locations to each of 791 SEQ 1D NO: 1-5,822. Methods A, B, and C were used to predict the polypeptides. Method A refers to a polypeptide obtained by using a software program called FASTY (available from http://fasta.bioch.virginia.edu) which selects a polypeptide based on a comparison of translated novel polynucleotide to known polypeptides (W.R. Pearson, Methods in

Enzymology, 183: 63-98 (1990), incorporated herein by reference). Method B refers to a polypeptide obtained by using a software program called GenScan for human/vertebrate sequences (available from Stanford University, Office of Technology Licensing) that predicts the polypeptide based on a probabilistic model of gene structure/compositional properties (C. Burge and S. Karlin, J. Mol. Biol., 268: 78-94 (1997), incorporated herein by reference). Method C refers to a polypeptide obtained by using a Hyseq proprietary software program that translates the novel polynucleotide and its complementary strand into six possible amino acid sequences (forward and reverse frames) and chooses the polypeptide with the longest open reading frame.

Novel Nucleic Acid Sequences Obtained From Various Libraries

A plurality of novel nucleic acids were obtained from cDNA libraries prepared from various human tissues and in some cases isolated from a genomic library derived from human chromosome using standard PCR, SBH sequence signature analysis and Sanger sequencing techniques. The inserts of the library were amplified with PCR using primers specific for the vector sequences which flank the inserts. Clones from cDNA libraries were spotted on nylon membrane filters and screened with oligonucleotide probes (*e.g.*, 7-mers) to obtain signature sequences. The clones were clustered into groups of similar or identical sequences. Representative clones were selected for sequencing.

In some cases, the 5' sequence of the amplified inserts was then deduced using a typical Sanger sequencing protocol. PCR products were purified and subjected to fluorescent dye terminator cycle sequencing. Single pass gel sequencing was done using a 377 Applied Biosystems (ABI) sequencer to obtain the novel nucleic acid sequences. In some cases RACE (Random Amplification of cDNA Ends) was performed to further extend the sequence in the 5' direction.

Novel Contigs

The contigs or the nucleic acids of the present invention, designated as 791 SEQ ID NO: 1-5,822 were assembled using an EST sequence from Hyseq's database as a seed. A recursive algorithm was used to extend the seed EST into an extended assemblage, by pulling additional sequences from different databases (*e.g.*, Hyseq's database containing EST sequences, dbEST version 115, gb pri 115, and UniGene version 103, and exons from public domain genomic sequences predicted by GenScan) that belong to this

assemblage. The algorithm terminated when there were no additional sequences from the databases that will extend the assemblage. Further, the inclusion of component sequences into the assemblage was based on a BLASTN hit to the extending assemblage with BLAST score greater than 300 and percent identity greater than 95%.

The nearest neighbor result for the assembled contig was obtained by a FASTA version 3 search against Genpept release 115, using FASTXY algorithm. FASTXY is an improved version of FASTA alignment which allows in-codon frame shifts. The nearest neighbor result showed the closest homologue for each assemblage from Genpept (and contains the translated amino acid sequences for which the assemblage encodes

6.13 **The 792 Gene Family**

The novel predicted polypeptides (including proteins) encoded by the novel polynucleotides (792 SEQ ID NO: 1-8,502) of the present invention, and their corresponding nucleotide locations to each of 792 SEQ ID NO: 1-8,502 were determined. Methods A, B, and C were used to predict the polypeptides. Method A refers to a polypeptide obtained by using a software program called FASTY (available from http://fasta.bioch.virginia.edu) which selects a polypeptide based on a comparison of translated novel polynucleotide to known polypeptides (W.R. Pearson, Methods in Enzymology, 183: 63-98 (1990), incorporated herein by reference). Method B refers to a polypeptide obtained by using a software program called GenScan for human/vertebrate sequences (available from Stanford University, Office of Technology Licensing) that predicts the polypeptide based on a probabilistic model of gene structure/compositional properties (C. Burge and S. Karlin, J. Mol. Biol., 268: 78-94 (1997), incorporated herein by reference). Method C refers to a polypeptide obtained by using a Hyseq proprietary software program that translates the novel polynucleotide and its complementary strand into six possible amino acid sequences (forward and reverse frames) and chooses the polypeptide with the longest open reading frame.

The isolated polypeptides of the invention include, but are not limited to, a polypeptide comprising any of the amino acid sequences predicted as described above or from six frame translations of 792 SEQ ID NO: 1-8,502; or the corresponding full length or mature protein. One of skill in the art could determine the corresponding amino acid

sequence using techniques well known in the art to translate and analyze all possible six frames. Polypeptides of the invention also include polypeptides with biological activity that are encoded by (a) any of the polynucleotides having a nucleotide sequence set forth in the 792 SEQ ID NO: 1-8,502; or (b) polynucleotides that hybridize to the complement of the polynucleotides of (a) under stringent hybridization conditions. Biologically or immunologically active variants of any of the polypeptide sequences or from six frame translations of 792 SEQ ID NO: 1-8,502, and "substantial equivalents" thereof (*e.g.,* with at least about 65%, 70%, 75 %, 80%, 85 %, 90%, 95 %, 98% or 99% amino acid sequence identity) that preferably retain biological activity are also contemplated. The polypeptides of the invention may be wholly or partially chemically synthesized but are preferably produced by recombinant means using the genetically engineered cells (e.g. host cells) of the invention.

Novel Nucleic Acid Sequences Obtained From Various Libraries

A plurality of novel nucleic acids were obtained from cDNA libraries prepared from various human tissues and in some cases isolated from a genomic library derived from human chromosome using standard PCR, SBH sequence signature analysis and Sanger sequencing techniques. The inserts of the library were amplified with PCR using primers specific for the vector sequences which flank the inserts. Clones from cDNA libraries were spotted on nylon membrane filters and screened with oligonucleotide probes (*e.g.,* 7-mers) to obtain signature sequences. The clones were clustered into groups of similar or identical sequences. Representative clones were selected for sequencing.

In some cases, the 5' sequence of the amplified inserts was then deduced using a typical Sanger sequencing protocol. PCR products were purified and subjected to fluorescent dye terminator cycle sequencing. Single pass gel sequencing was done using a 377 Applied Biosystems (ABI) sequencer to obtain the novel nucleic acid sequences. In some cases RACE (Random Amplification of cDNA Ends) was performed to further extend the sequence in the 5' direction.

Novel Contigs

The contigs or the nucleic acids of the present invention, designated as 792 SEQ ID NO: 1-8,502 were assembled using an EST sequence from Hyseq's database as a seed. A recursive algorithm was used to extend the seed EST into an extended assemblage, by

pulling additional sequences from different databases (*e.g.*, Hyseq's database containing EST sequences, dbEST version 115, gb pri 115, and UniGene version 103, and exons from public domain genomic sequences predicted by GenScan) that belong to this assemblage. The algorithm terminated when there were no additional sequences from the databases that will extend the assemblage. Further, the inclusion of component sequences into the assemblage was based on a BLASTN hit to the extending assemblage with BLAST score greater than 300 and percent identity greater than 95%.

The nearest neighbor result for the assembled contig was obtained by a FASTA version 3 search against Genpept release 115, using FASTXY algorithm. FASTXY is an improved version of FASTA alignment which allows in-codon frame shifts. The nearest neighbor result showed the closest homologue for each assemblage from Genpept (and contains the translated amino acid sequences for which the assemblage encodes).